Data Privacy Project

I inspired to do this project from my research about Datavant and what exactly a Data Scientist – Data Privacy role is doing.

For this project I used a data set from Kaggle named ‘’Heart_disease_uci ‘’ and contains 920 rows and 16 columns, including patient demographics (Age, Sex), clinical features (cp (chest pain type), chol), and a target variable (num) indicating the presence and danger of heart disease.

Our primary goal was to anonymize the dataset in order to minimize the risk of re-identifying individuals while preserving the dataset's value for analysis and to be more familiar with terminologies like QIDs (Quasi-Identifiers), k-Anonymity, l-Diversity, t-Closeness and industry-standard privacy-preserving techniques.

As 1st step we selected the columns most relevant to privacy risk: age, sex, and cp. These are known as quasi-identifiers, and their main characteristic is that as individual attributes may not be unique on their own but can become identified when combined. Also, we could have included the dataset column, which specifies the source hospital (e.g., Cleveland, Hungary), but in our case, we considered it less sensitive.

This step is critical because it determines which columns we use to group individuals when evaluating re-identification risk. By identifying and combining quasi-identifiers into a QID (Quasi-Identifier Group), we can begin to understand how many people share the same profile, and eventually, how easy it would be for someone to be targeted out in the dataset.

Next we focused on preparing the dataset for privacy analysis by keeping only the relevant quasi-identifiers (age, sex, cp) and the sensitive attribute or target column (num). We then proceeded to clean the data by removing or correcting any missing values to ensure consistency. Also, to avoid inconsistencies in categorical variables, we standardized fields like sex and cp by converting them to lowercase and ensuring they were stored as string types. Finally, this stage is crucial as it ensures the dataset is in a consistent and structured format, enabling accurate grouping, comparison, and risk assessment in the next steps of anonymization.

The goal of this step was to make the quasi-identifiers (age, sex, cp) less specific in order to reduce the risk of individual identification and help achieve k-anonymity. To do this, we grouped age into broad age bands (e.g., 0–40, 41–50, etc.) instead of using exact ages, generalized the cp column by reducing the 4 original categories to 3, combining similar types into broader groups such as "angina" and "other", simplified the sensitive attribute by converting the original 5 levels into a binary format: 0 for no disease and 1 for any presence of heart disease and finally, we created a new column called QID by combining the generalized age, sex, and cp values. This unified identifier was used to group individuals with similar characteristics for further privacy analysis.

Then, we evaluated how many individuals share the same QID by calculating the size of each group. This allowed us to measure k-anonymity, which refers to the number of people who share the same combination of quasi-identifiers. If a person belongs to a QID group of size 1, they are uniquely identifiable and that causes a serious privacy risk. During our analysis, we found that the smallest

QID group had a size of 7, indicating that every individual in the dataset shared their profile with at least 6 others. This confirmed that the dataset satisfies k ≥ 2-anonymity, which is an important foundation for privacy protection.

While k-anonymity helps protect who someone is, l-diversity ensures we also protect what someone has.

In this step, we evaluated whether each QID group contained more than one distinct value for the new sensitive attribute (num_gen).

A QID group is considered l-diverse if it includes at least 1 distinct diagnosis. A QID group is considered l-diverse if it includes at least l distinct diagnoses. Without l-diversity, attackers who know a person's quasi-identifiers could still conclude to diagnosis with high certainty, even if they can't identify the exact record.

We calculated the number of unique diagnosis values within each QID group, and we found out that 907 records have at least two different diagnoses present and only 13 share the same diagnosis.

This analysis showed that most of records (98.6%) were l-diverse, satisfying l ≥ 2. However, 13 records were found in QID groups with l = 1, creating a risk of attribute disclosure.

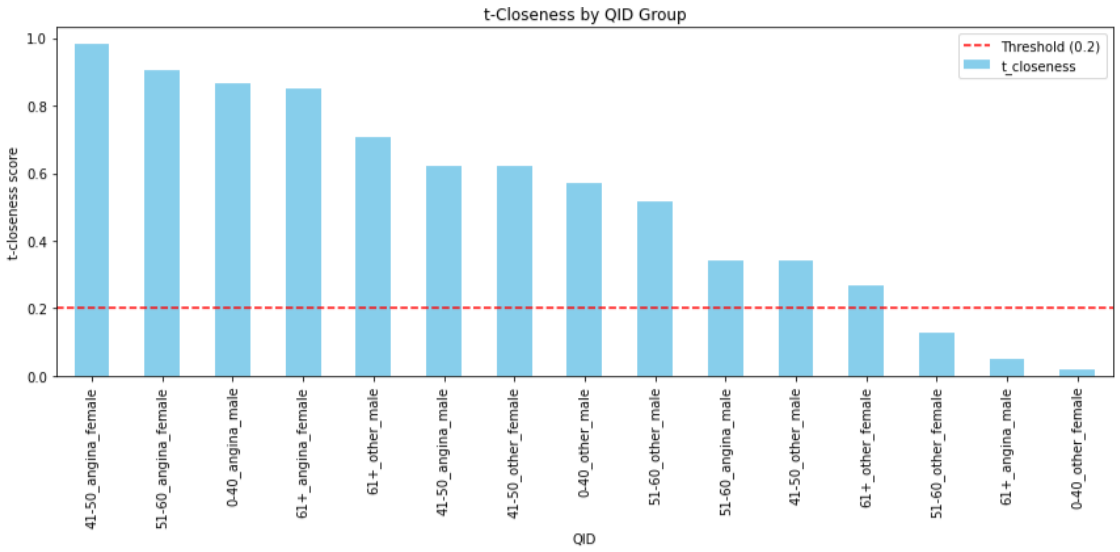
Given that these 13 instances represented only 1.4% of the dataset and came from highly specific groups, we chose to remove them to ensure that the entire dataset satisfies the l-diversity requirement.

Finally, after securing k-anonymity (protecting identity) and l-diversity (protecting exact diagnosis), we moved on to t-closeness, which addresses a more subtle privacy risk trying to answer to ''What if a QID group's diagnosis distribution is very different from the overall population?''

Even if a group is l-diverse, if all values are skewed in one direction, attackers could still make accurate inferences. t-Closeness ensures that each group's sensitive attribute distribution is close to the overall distribution, limiting the risk of attribute inference. To achieve that we used the sum of absolute differences between the diagnosis distribution within each QID group and the global distribution of diagnoses (num_gen). This is a practical approximation of the Earth Mover's Distance (EMD) used in t-closeness calculations.

We calculated the overall diagnosis distribution and the diagnosis distribution within each QID group, after we took the absolute difference between the two distributions for each group and eventually we created a t-closeness table with the summed the differences and t-closeness score per QID group.

Each QID group was assigned a t-closeness score, and high-risk groups (with scores > 0.4) were flagged for further action. This final step allowed us to evaluate the distributional privacy risk and ensure that even statistical inference would be difficult, completing our anonymization pipeline.


t-Closeness by QID Group
Threshold (0.2)
t_closeness
t-closeness score
1.0
0.8
0.6
0.4
0.2
0.0
41-50_angina_female
51-60_angina_female
0-40_angina_male
61+_angina_female
61+_other_male
41-50_angina_male
41-50_other_female
0-40_other_male
51-60_other_male
51-60_angina_male
41-50_other_male
61+_other_female
51-60_other_female
61+_angina_male
0-40_other_female
QID