Project-2: Privacy vs. Utility

This report provides an overview of the anonymization project, including details about the choice of k and ℓ values, the definition of quasi-identifiers (QID) and the sensitive attribute, the method used to measure utility, and the results demonstrating how utility was impacted by changes in k and ℓ.

1. Choice of k and ℓ Values

We employed different combinations of k and ℓ values to assess the impact on data privacy and utility. Specifically, the following values were used:

- k-values = [15, 20, 50]: These represent different levels of k-anonymity, where a higher k indicates a higher level of anonymity.
- ℓ-values = [3, 6, 15]: These represent different levels of ℓ-diversity, where a higher ℓ implies a higher level of diversity within sensitive attribute values.

2. QID and Sensitive Attribute

In this project, the QID were defined as a combination of the following attributes:

- Gender: Represents the gender of passengers.
- Airport Continent: Indicates the continent where the departure airport is located.
- Departure Date: Represents the date of departure, which includes month and year.

These attributes were selected because they are non-sensitive and commonly found in flight datasets. By considering these attributes as QID, the goal was to ensure that each record in the dataset becomes indistinguishable from at least $k-1$ other records based on these attributes, achieving k-anonymity.

The sensitive attribute chosen for this project is **Flight Status**. This attribute indicates the status of a flight (e.g., "On Time," "Delayed," or "Cancelled"). The choice of this attribute as sensitive is based on the need to protect information related to flight delays and cancellations, which can be sensitive in certain contexts.

3. Utility Measurement

Utility was measured in this project using KL-divergence, which is a statistical measure that quantifies the divergence between probability distributions of the original and anonymized data for a specified sensitive attribute (Flight Status in this case). The KL-divergence metric allows us to assess how closely the anonymized data distribution matches the original data distribution for the sensitive attribute.

4. Impact of k and ℓ on Utility

- Lower values of k and ℓ (e.g., k=15 and ℓ=3) tend to preserve higher levels of utility. In this scenario, the anonymized data retains a significant degree of similarity with the original data.

- As k increases (e.g., k=20 and k=50), the utility tends to decrease. Higher values of k lead to a more generalized and anonymized dataset, which may result in a larger discrepancy between the original and anonymized data distributions.

- Similarly, increasing ℓ (e.g., from ℓ=3 to ℓ=6 and ℓ=15) also correlates with a decrease in utility. Higher ℓ values demand more diversity within the sensitive attribute values, which can lead to further distortion of the data and reduced utility.

These results highlight the trade-off between data privacy and utility. Higher levels of anonymity (achieved through larger k and ℓ values) enhance privacy but come at the cost of reduced data utility. The choice of k and ℓ values should be driven by the specific privacy requirements and the acceptable level of utility for the intended use case.