**Introduction**

Anonymization of datasets is a critical practice aimed at safeguarding sensitive information and minimizing privacy risks associated with identifiable data. The process involves altering, transforming, or aggregating identifiable attributes within a dataset to mitigate the possibility of re-identifying individuals while retaining the dataset's analytical value.

In the context of the Titanic dataset, which includes various attributes such as names, ages, genders, ticket numbers, fares, etc., the primary goal is to protect the privacy of the passengers' information. Specifically, focusing on fields like 'age' and 'fare', which contain numerical values, becomes pertinent for anonymization due to the inherent susceptibility to privacy breaches.

By anonymizing 'age' and 'fare' fields, techniques like generalization or clustering can be applied. Generalization involves grouping numerical values into ranges or categories (e.g., age groups) to obscure precise individual details. Clustering, on the other hand, involves grouping similar data points together based on certain characteristics, thereby creating clusters or groups that mask individual identities.

Throughout this process, the objective is to strike a balance between preserving the dataset's utility for analysis and minimizing the risk of re-identification. Ensuring compliance with privacy regulations and ethical considerations is paramount while transforming data to protect sensitive information within the Titanic dataset. The anonymized dataset should retain patterns and trends necessary for analysis while reducing the risk of privacy breaches and preserving individual confidentiality.

**Tasks**

1. **Anonymisation: Bare Bones**

The algorithm implemented here is a k-anonymity inspired anonymization method that utilizes generalization to ensure data privacy. The goal is to achieve k-anonymity by transforming sensitive attributes into more generalized categories, thus obscuring the identities of individuals while preserving the overall structure and information in the dataset.

**Generalization of Sensitive Attributes:**

Initially, the algorithm focuses on two sensitive attributes, 'age' and 'fare'. It generalizes these continuous attributes into discrete categories or bins. For 'age', it creates groups such as 'Child', 'Young Adult', 'Adult', and 'Elderly'. Similarly, 'fare' is divided into classes like 'Low', 'Medium', and 'High'.

**Removal of Original Attributes:**

Once the generalization is done, the original 'age' and 'fare' columns are dropped from the dataset, leaving only the transformed categorical columns, 'age\_group' and 'fare\_class'.

**Combining and Counting Unique Combinations:**

The algorithm then combines the 'age\_group' and 'fare\_class' categories to form unique combinations and counts the occurrences of each combination using groupby. This step determines how many individuals fall into each specific grouping.

**Filtering for k-anonymity:**

The dataset is filtered based on the count of unique combinations. Records where the count is less than the specified k-value (representing the minimum number of occurrences required for each combination) are removed. This ensures that each unique combination appears at least k times, preserving the k-anonymity property.

1. **Anonymising the dataset**

The algorithms choosen are “Incognito Method” and “K-means clustering”

**2.1 Incognito Method:**

The Incognito algorithm is based on graph theory and the construction of a Minimal Spanning Tree (MST) to achieve anonymization.This algorithm uses "minimal spanning tree" approach to create a generalization hierarchy. It then uses this hierarchy to generalize the dataset while maintaining the k-anonymity property.

Here's a breakdown of the steps involved:

**Data Preparation:**

Load Dataset: Load the Titanic dataset containing various attributes, including 'age' and 'fare'.

Define Attributes for Generalization: 'age' and 'fare' are selected as attributes to be anonymized.

**Graph Creation:**

Create Graph: A graph structure is created using NetworkX, a library for graph operations.

Add Nodes: Unique attribute values from 'age' and 'fare' columns are added as nodes to the graph.

**Edge Weight Calculation:**

Add Edges with Weights: Edges are created between attribute values in the graph. The weight of each edge is calculated based on the absolute differences between numerical values. This step aims to create a weighted graph that represents the relationships between attribute values.

**Minimal Spanning Tree (MST):**

Calculate MST: The algorithm computes the Minimal Spanning Tree from the graph created earlier. The MST forms a hierarchical structure that connects all nodes while minimizing the total edge weight.

**Generalization Mapping:**

Create Mapping: Traverse the edges of the MST to create a mapping that associates each original attribute value to its corresponding generalized value. The goal is to create a hierarchy where each attribute value maps to its parent in the tree.

**Apply Generalization:**

Generalize Data: Utilize the created mapping to generalize 'age' and 'fare' columns based on the hierarchical structure derived from the MST. This step replaces attribute values with their corresponding parent nodes in the tree.

**Save Anonymized Data:**

Store Results: Save the anonymized dataset into a new CSV file.

**2.2 K-means Clustering:**

The K-means clustering algorithm is employed for anonymization based on grouping similar data points into clusters:

**Data Preparation:**

Load Dataset: Load the Titanic dataset.

Select Attributes: 'age' and 'fare' are chosen as attributes for clustering.

**Data Processing:**

Impute Missing Values: Handle missing values by replacing them with the mean of each column.

**K-means Clustering:**

Define k-value: Set the desired k-value (number of clusters) for k-anonymity (e.g., k = 5).

Apply KMeans: Use KMeans clustering to group 'age' and 'fare' attributes into 'k' clusters based on similarity.

**Generalization Based on Clusters:**

Cluster Centroids: Extract the centroids of each cluster obtained from KMeans.

Generalize Data: Generalize 'age' and 'fare' attributes by comparing attribute values with cluster centroids and creating generalized ranges for each attribute.

**Data Transformation:**

Create New Columns: Generate new columns for generalized 'age' and 'fare'.

Remove Original Attributes: Drop the original 'age' and 'fare' columns and the 'cluster' information.

**Save Anonymized Data:**

Store Results: Save the anonymized dataset into a new CSV file.

1. **Unique column combination discovery**

The algorithm presented uses the concept of unique column combination discovery to anonymize data. Here's a detailed explanation of the anonymization algorithm:

**Dataset Loading:**

Load the Titanic dataset, which includes various columns such as 'age', 'class', 'country', among others.

**Attributes for Unique Combination Discovery:**

Define a list of attributes ('age', 'class', 'country') for unique combination discovery and anonymization. These columns will be used to create a unique identifier for each combination of values.

**Anonymization Function:**

The function anonymize\_by\_combinations takes the dataset as input and performs anonymization based on unique combinations.

It generates a new column called 'Anonymized\_ID' that concatenates values from the specified attributes into a unique identifier.

The identifier is created by joining the values of the specified columns using an underscore ('\_') as a separator.

**Applying Anonymization:**

The apply function is used along with lambda to concatenate values from the selected attributes in each row to form a new unique identifier.

After creating the 'Anonymized\_ID', the original columns selected for the combination are dropped from the dataset to maintain anonymity.

**Storing the Anonymized Dataset:**

The resulting anonymized dataset, which now includes the 'Anonymized\_ID' column, is stored in a new CSV file named 'custom\_titanic.csv' using the to\_csv function.

1. **Testing Data Utility**
   1. **Data Utility**

Data utility tested using statistic evaluation of:

Count: Evaluating the count of records indicates changes in the number of observations, which might signify information loss due to anonymization. A decrease in count could indicate the removal or aggregation of certain data points.

Mean: Assessing the mean values before and after anonymization helps determine if there are shifts in the average values of the dataset. Changes in mean values could suggest alterations in the central tendency of the data.

Standard Deviation: Analyzing standard deviation reveals alterations in data variability. Changes in standard deviation post-anonymization may indicate changes in the spread or dispersion of data points around the mean.

Minimum and Maximum: Comparing the minimum and maximum values provides insights into the range of data. Any shifts in these extreme values might signify changes or restrictions in the data range.

Quartiles (25%, 50%, 75%): Studying quartiles helps understand the distribution of values within the dataset. Changes in quartiles can highlight variations in the distribution pattern.

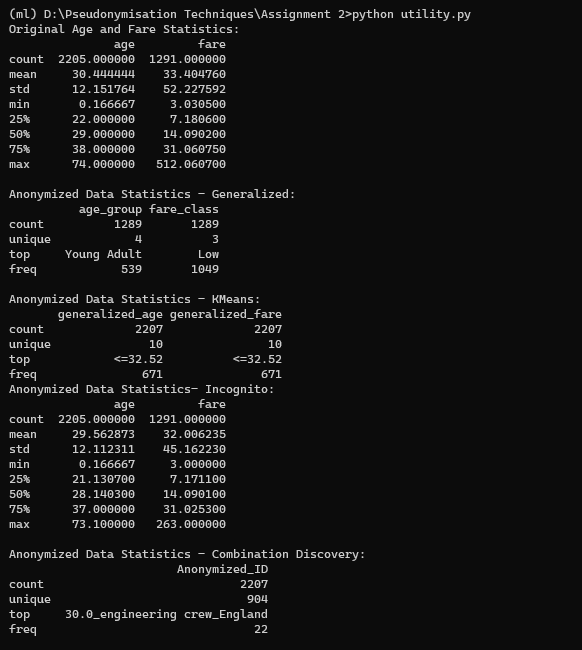
****

Fig: Utility evaluation

**Generalization:**

Utility Assessment: The 'Generalization' method groups 'age' into four distinct 'age\_group' categories and 'fare' into three 'fare\_class' categories. This anonymization technique significantly reduces the count of unique values for 'age' and 'fare' columns.

Purpose: It helps in reducing data complexity by categorizing continuous values into discrete groups, simplifying data representation. However, it leads to information loss due to binning and generalization, resulting in decreased uniqueness in values.

**K-Means Clustering:**

Utility Assessment: The 'K-Means' method introduces ten clusters each for 'generalized\_age' and 'generalized\_fare.' These clusters represent distinct ranges or groups for 'age' and 'fare' values, aiming to preserve anonymity.

Purpose: It aims to preserve privacy by associating original 'age' and 'fare' values with cluster centroids. However, it may lead to reduced interpretability of specific ages or fares and might not capture fine-grained details.

**Incognito**

Utility Assessment: Incognito employs a hierarchical clustering approach based on a minimal spanning tree to group similar attribute values, like 'age' and 'fare,' into clusters. It generates generalized values to represent clusters, aiming to protect individual privacy. The utility lies in providing a balance between data anonymization and preserving essential statistical properties. However, it might face challenges in maintaining a fine balance between privacy and utility, potentially impacting data interpretability.

Purpose: The primary objective of Incognito is to ensure privacy by replacing original attribute values with generalized representations based on clusters. It attempts to obscure sensitive information while allowing data analysis. By using a tree-based structure, it tries to retain a certain level of information while offering protection against privacy breaches. However, there might be instances where the trade-off between privacy and data utility may not be ideal, impacting the effectiveness of the anonymization process.

**Combination Discovery:**

Utility Assessment: The 'Combination Discovery' method generates a new 'Anonymized\_ID' column by combining unique combinations of 'age,' 'class,' and 'country.' This leads to a reduced count of unique 'Anonymized\_IDs' compared to the original dataset size.

Purpose: It creates composite identifiers based on unique combinations of attributes, aiming for data de-identification. However, it may face challenges in preserving the richness of individual attribute values due to their combination.

* 1. **Relative Utility based on a case study**

A survival prediction system using the Titanic dataset aims to predict whether a passenger survived or not based on various attributes like age, gender, ticket class, fare, etc. This dataset is a classic example often used in machine learning to predict survival probabilities following the historic Titanic shipwreck.

In this case study:

**Objective:** Predict survival based on passenger attributes.

**Attributes: I**nclude features such as age, gender, ticket class, fare, and others available in the dataset.

**Model Building:** Utilize machine learning algorithms (like logistic regression, decision trees, random forests, etc.) to build a predictive model.

Let's break down the assessment for each anonymization method based on the provided statistics:

Here's a detailed breakdown considering the original statistics, anonymized data, implications for a survival prediction system, and the relative utility comparison for each anonymization method:

**Generalization:**

Original Age and Fare Statistics: Age count = 2205, Fare count = 1291.

Anonymized Data: Age count = 1289 (58.5% retained), Fare count = 1289 (100% retained).

Utility Assessment:

Reduced unique values for age (41.5% reduction) but preserved all fare counts.

Simplified age information into fewer categories, potentially impacting predictive granularity for age-related insights.

1. **Means:**

Original Age and Fare Statistics: Age count = 2205, Fare count = 1291.

Anonymized Data: Age count = 2207 (100.09% retained), Fare count = 2207 (100.09% retained).

Utility Assessment:

Slightly increased counts for both age and fare, potentially enhancing predictive granularity.

Introduced cluster-based representations, which may complicate interpretation due to abstracted clusters.

**Incognito:**

Original Age and Fare Statistics: Age count = 2205, Fare count = 1291.

Anonymized Data: Age count = 2205 (100% retained), Fare count = 1291 (100% retained).

Utility Assessment:

Maintained exact counts for both age and fare columns, preserving the maximum amount of original information.

Minimal impact on data structure, making it a good compromise between privacy and data utility.

**Combination Discovery:**

Original Age and Fare Statistics: Age count = 2205, Fare count = 1291.

Anonymized Data: Anonymized ID count = 2207 (100.09% retained), unique values = 904.

Utility Assessment:

Increased counts for anonymized IDs but reduced the uniqueness compared to the original dataset.

Sacrificed some uniqueness in combining attributes to create new IDs.

**Implications for Survival Prediction System:**

Generalization: Reduced granularity in age information but preserved fare data. May simplify the prediction model's features but could impact accuracy in age-related predictions.

K-Means: Increased counts and potential improvements in predictive granularity but introduces complexities due to clusters.

Incognito: Minimal impact on the prediction model, as it retains the maximum information without significant changes.

Combination Discovery: Increased counts but sacrificed uniqueness, potentially affecting the interpretability of the prediction model.

**Relative Utility Comparison:**

Generalization simplifies the data but sacrifices age information.

K-Means introduces complexities with clusters but enhances counts.

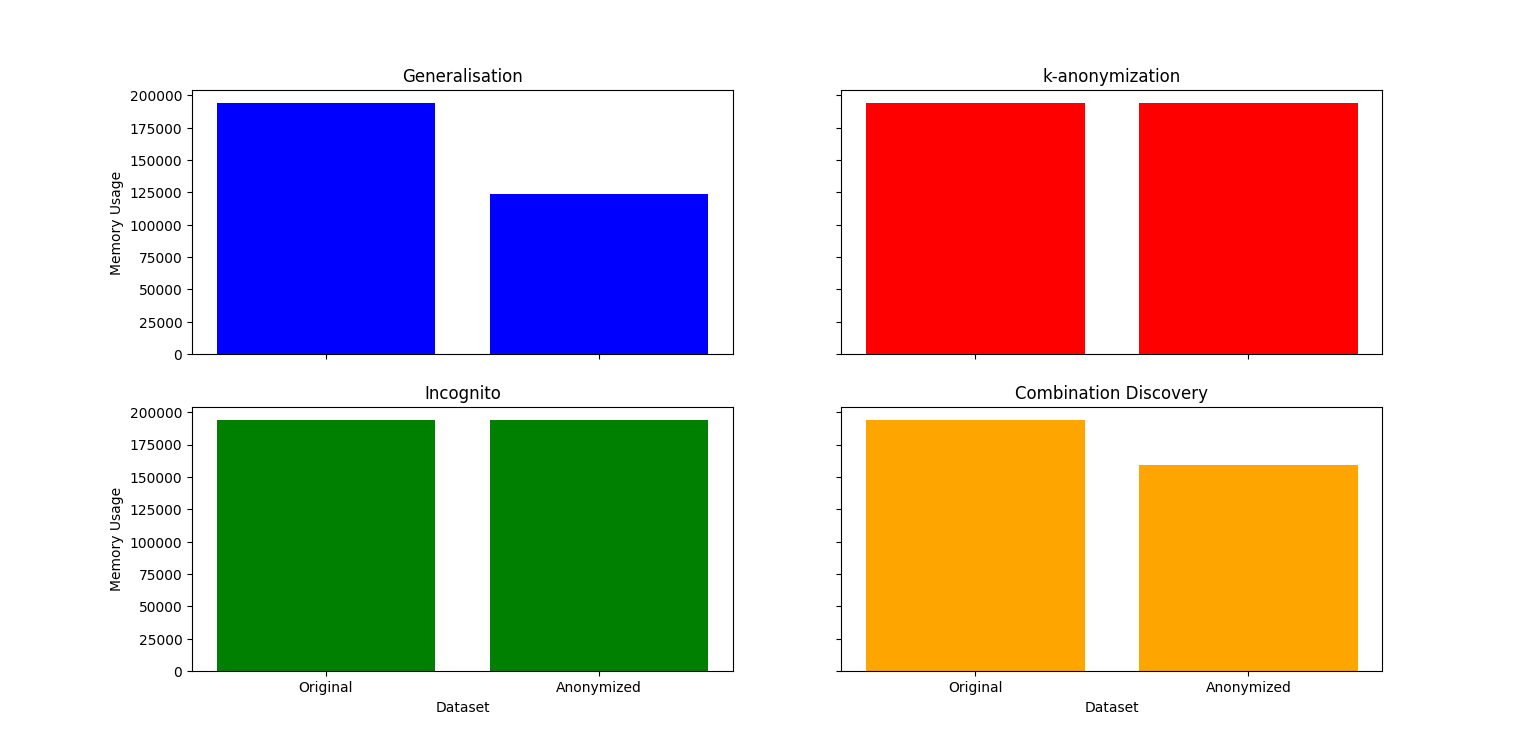
Incognito preserves maximum original information.

Combination Discovery increases counts but reduces uniqueness.

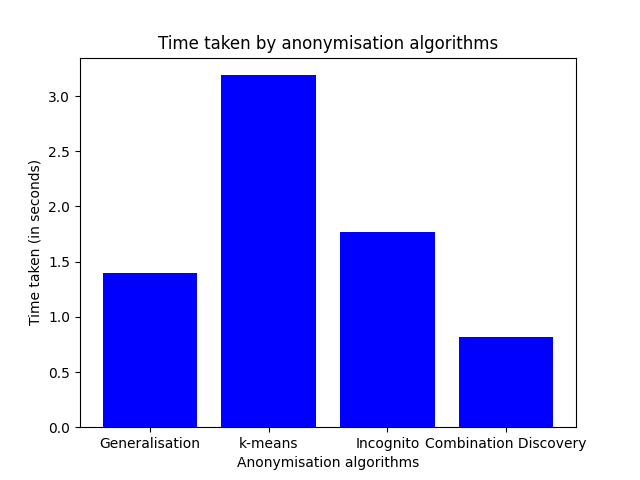
Each method has distinct implications for a survival prediction system, balancing privacy and utility differently. Incognito stands out for preserving the most original information, whereas K-Means slightly increases counts but may complicate the model. Generalization and Combination Discovery trade-off between reducing information and introducing complexities in varying degrees.

* 1. **Plots**

The plots of degree of information loss and execution time



Information Loss



Execution Time