**1. ABSTRACT**

Myriads of privacy preserving algorithms are present but very few aim to take into consideration the privacy for multiple sensitive attributes. Many pre-existing algorithms aim to segregate the table into sensitive and non-sensitive tables or completely mask or generalize the information. This paper combines the best of both worlds and develops an innovate algorithm termed as incremental diversity.

Incremental diversity algorithm is able to diversify the data for multiple sensitive attributes and also produce lesser quantity of residue records. Incremental diversity mainly chooses a primary sensitive attribute having a certain set of favorable characteristics such the presence of a greater number of unique values as compared to other sensitive attributes, many parents in its semantic hierarchical tree, and also more varying frequency for each sensitive value in the dataset.

The inference obtained after careful examination of the graphs is that incremental diversity has much better time complexity, generates lesser number of residue records for a given value of k, but it comes at the cost of having lesser diversity than (l, e) diversity.

The algorithm can be suitably adjusted to balance the trade-off between diversity and residue records produced for a given dataset depending on the algorithm’s use-case and the degree of sensitivity that is desired. Privacy can be easily strengthened by choosing the most suitable primary sensitive attribute of the microdata.

**2. ALGORITHM**

*Input*: Microdata Table (M1), k, l, e, Total no. of records (N)

*Output*: QIT (q1, q2, q3, …, qn), ST (s1, s2, s3, …, sn), (l, e) Masked Microdata Table (MMT)

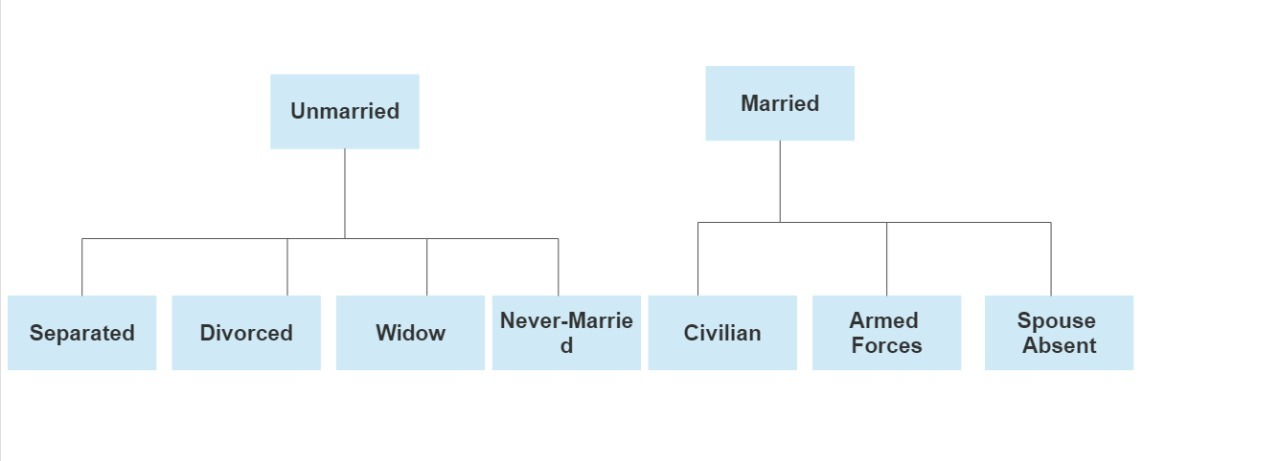
*Steps:*

1. Standardization of Microdata and storage in a nested dictionary format.
2. Assignment of the Group ID (or more specifically, the equivalence class number) to each record with the formula given below.
3. Diversification of records by any one of the five different conditions, on the basis of which records are removed from the microdata table, added to the temporary dictionary, and added back to the modified microdata table (discussed in Tables and Formulae Section).
4. Creation of a separate dictionary (called the temporary dictionary) to store the records that do not fit in the selected equivalence class of the microdata table on the basis of the conditions associated.
5. Picking and placing of the records from the temporary dictionary into the new microdata dictionary based on the pre-selected parameters along with the extra condition that the size of the existing equivalence class should be strictly lesser than k (k-anonymity [30]).
6. Increase in the diversity of Secondary, Tertiary and Quaternary Sensitive Attributes by swapping their sensitive values which are repeating in a particular equivalence class with non-repeating values from the temporary dictionary.
7. Segregation of the modified microdata table into Quasi-Identifier Table (QIT) and Sensitive Table (ST).
8. Masking and generalisation of the attributes with ease & efficiency (formulas discussed in Tables & Formulae Section).

**3. TABLES AND FORMULAE**

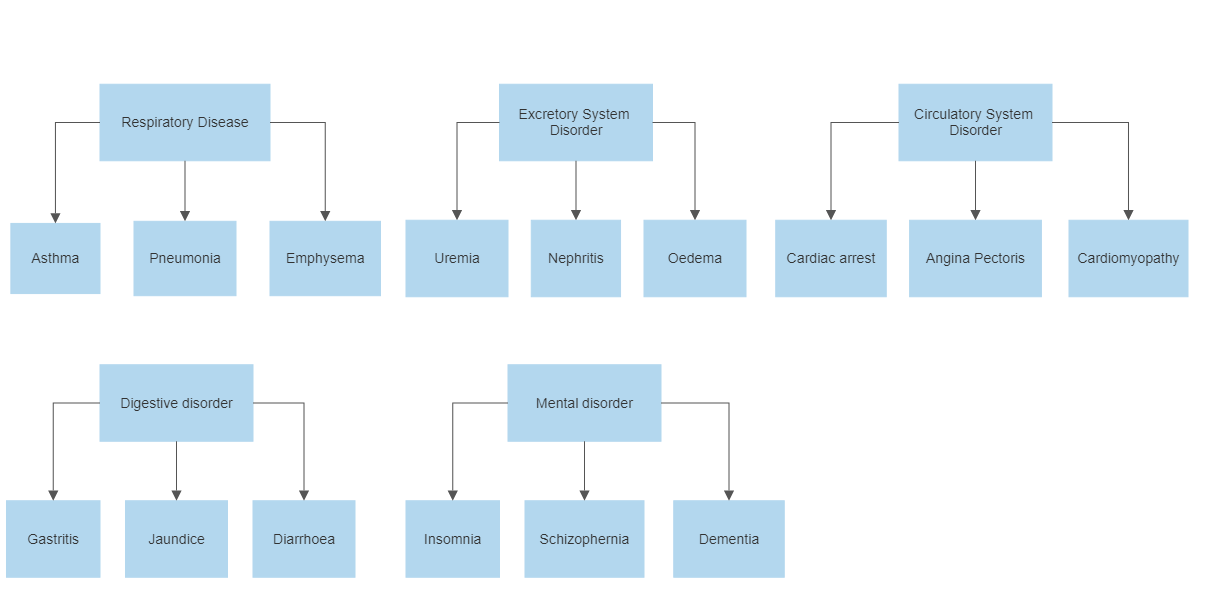
The five diversification conditions are:

1. Based on the unique appearance of a primary sensitive attribute with few parents in its semantic hierarchical tree (Eg: Marital Status): In a particular equivalence class, a value of the chosen primary sensitive attribute appears only once, i.e., the sensitive value is unique for a specific equivalence class.
2. Based on the unique appearance of a parent in the semantic hierarchical tree consisting of a fewer number of parents for the primary sensitive attribute: It is similar to the former algorithm with the key difference being that instead of the sensitive value, the parent of the chosen primary sensitive attribute appears only once in the group, i.e., it is unique for a specific equivalence class. Example: If a record in an equivalence class has the value of “Widow” for the primary sensitive attribute “Marital Status”, the parent of “Widow” being “Unmarried”, no other record containing Marital Status value as Separated, Divorced or Never-Married is allowed to be added into the equivalence class.
3. Based on the appearance of a parent in the semantic hierarchical tree consisting of a fewer number of parents for the primary sensitive attribute at most twice: The parent of the chosen primary sensitive attribute is allowed to appear twice in a particular equivalence class.



**Figure 4.1: Semantic Hierarchical Tree for Marital Status**

1. Based on the unique appearance of a primary sensitive attribute with more varying frequency of each value in the dataset (Eg: Relationship): It is an offshoot of the first algorithm where only the primary sensitive attribute is changed from one having fewer parents in its sem­antic hierarchical tree (Marital Status) to the attribute having more variations in its values (Relationship).
2. Based on the unique appearance of a parent in the semantic hierarchical tree consisting of high number of parents for the primary sensitive attribute: A primary sensitive attribute with a higher no. of parents in its semantic hierarchical tree is chosen (Disease) and the parent uniquely appears in an equivalence class. For example, if in an equivalence class there is the presence of a disease for a particular record, let’s say, Cardiomyopathy, belonging to the Circulatory System Disorder Category (Parent), then no other record with the disease belonging to the same parent (Circulatory System Disorder) will be allowed to be placed in that equivalence class.



**Figure 4.2: Semantic Hierarchical Tree for Disease**

**Quasi Identifier Table (QIT) & Sensitive Table (ST)**

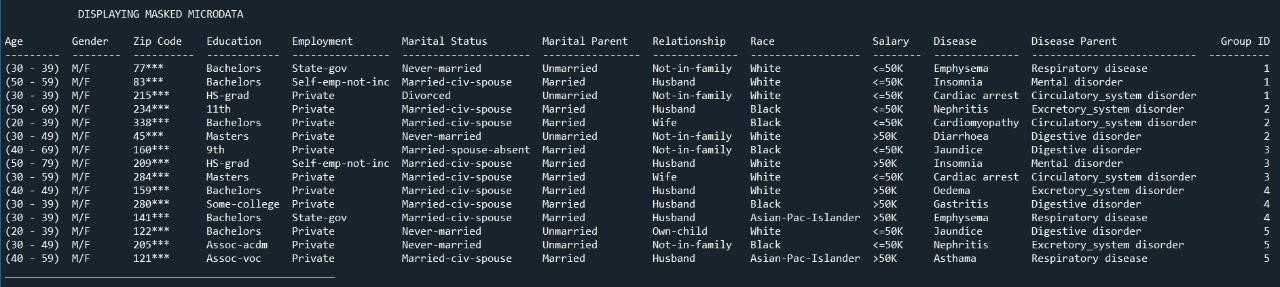
**Table 4.13: Quasi – identifier table**



**Table 4.14: Sensitive table (ST) Table 4.15: Count table**



**Table 4.16: Masked microdata table**



Formulae incorporated for masking and generalising the chosen sensitive attributes:

Generalising Age:

This a type of cyclic masking with step size of 3.

Generalising Gender: “M/F” is used.

Masking Zip Code: Only the last three digits are hidden with (\*). (Eg: 560\*\*\*)

Masking Employment, Race or Salary: Asterisk symbol (\*) is used.

The various performance parameters to evaluate and analyse the algorithm for varying values of the inputs (no. of records and k) is calculated with the following formulas:

Residue Percentage =

Diversity Percentage =

The diversity for each NMA is calculated as follows:

Diversity for each NMA =