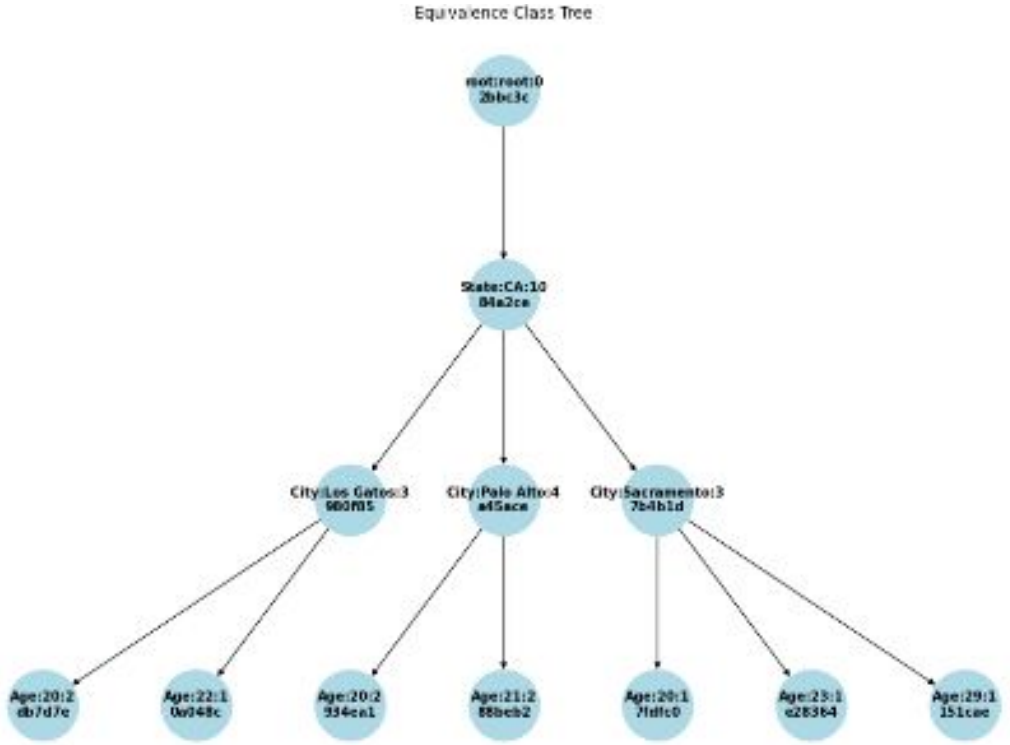
Equivalence Class Approach

Tree diagram from tree_example.ipynb



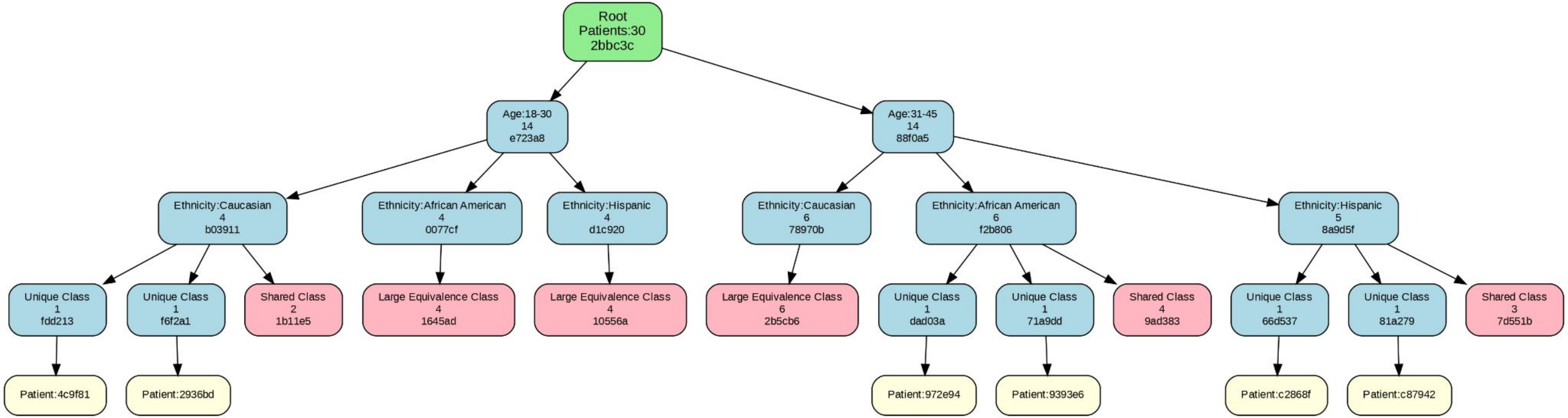
Equivalence Class Tree

Handling Division by Zero in the Uniqueness Ratio

Since we can't divide by zero, we need to adjust the calculation when the **bottom-layer sibling count is zero**. One common approach to handle this situation is:

- If the sibling count at the bottom layer (Cali Resident) is zero (i.e., the data point is unique at that layer), we can:
- **Skipping the calculation for those entries** (because the uniqueness is already absolute). — we need to keep track of uniqueness.
- **Assigning a predefined value** for cases where division by zero would occur (such as setting the uniqueness ratio to `-1` for absolute uniqueness).

Potential Equivalence Class Tree



Root
Patients:30
2bbc3c

Age:18-30
14
e723a8

Age:31-45
14
88f0a5

Ethnicity:Caucasian
4
b03911

Ethnicity:African American
4
0077cf

Ethnicity:Hispanic
4
d1c920

Ethnicity:Caucasian
6
78970b

Ethnicity:African American
6
f2b806

Ethnicity:Hispanic
5
8a9d5f

Unique Class
1
fdd213

Unique Class
1
f6f2a1

Shared Class
2
1b11e5

Large Equivalence Class
4
1645ad

Large Equivalence Class
4
10556a

Large Equivalence Class
6
2b5cb6

Unique Class
1
dad03a

Unique Class
1
71a9dd

Shared Class
4
9ad383

Unique Class
1
66d537

Unique Class
1
81a279

Shared Class
3
7d551b

Patient:4c9f81

Patient:2936bd

Patient:972e94

Patient:9393e6

Patient:c2868f

Patient:c87942

Potential Strategy

Generate Combinations: Test multiple combinations of selected rows (indirect identifiers).

Monitor Equivalence Classes: Track the number of unique equivalence classes for each combination.

Threshold: Set a threshold for how many equivalence classes are allowed. If there are too many, that combination is rejected.

Goal: Avoid combinations that could make patient identification too easy.

Optimized Potential Strategy

k-Anonymity: Ensure that each equivalence class has at least k individuals, preventing any small group from being uniquely identified.

l-Diversity: Guarantee that sensitive attributes within each equivalence class are diverse enough to prevent exposure of private data.

t-Closeness: Ensure that the distribution of sensitive attributes within an equivalence class is similar to the overall dataset, preserving privacy by reducing outliers.

Explanation of Techniques

k-Anonymity: — I think this approach is best

- **Definition**: Ensures that each individual is indistinguishable from at least `k-1` others in the dataset based on the selected indirect identifiers.
- **Example**: If k=5, every combination of indirect identifiers (e.g., ZIP code, gender, birth date) should appear in at least 5 rows.

l-Diversity:

- **Definition**: Extends k-anonymity by requiring that sensitive attributes (e.g., disease, income) within each equivalence class are sufficiently diverse.
- **Example**: In a medical dataset, if 5 people share the same quasi-identifiers, their disease attribute must be diverse (e.g., at least two different diseases are present).

t-Closeness:

- **Definition**: Requires that the distribution of sensitive attributes within each equivalence class is close to the distribution of those attributes in the entire dataset.
- **Example**: If 20% of the dataset has a certain disease, each equivalence class should have approximately 20% of individuals with that disease, preventing attackers from using distribution patterns to infer information.

Need to Check:

- If statistical significance is maintained
- If this reduces privacy testing time
- If these privacy checks are already being done
- If we want to let the client know once tree populates to not select these columns or to select particular columns for anonymity