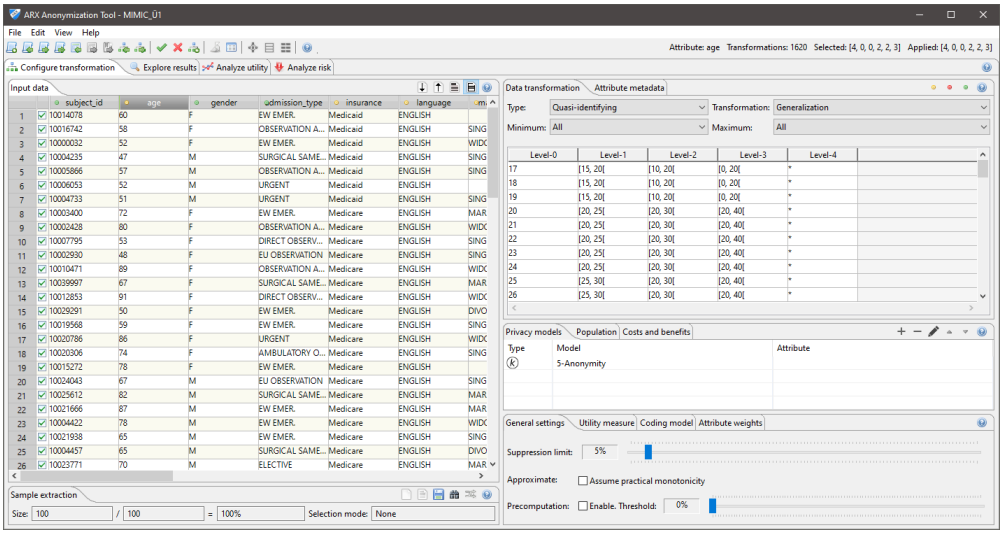
ARX MIMIC Übung

Basiert auf der ARX Übung von Dr. Fabian Prasser. Musterlösung

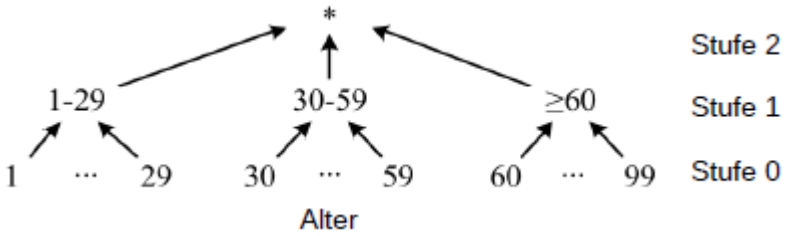
Aufgabe 1)

1. Öffnen Sie ARX und erstellen Sie ein neues Projekt. Importieren Sie den Datensatz “MIMIC IV Demo full Ü3.csv”. .
2. Wählen Sie die Spalten “age”, “insurance”, “language”, “martial_status”, “race” und “weight” nacheinander aus und wählen Sie jeweils den Attributtyp “Quasi-identifying”. Importieren Sie nun unter „File“ → „Import hierarchy” die entsprechende Hierachie für jedes Attribut das Sie ausgewählt haben.



Hintergrund

- Generalisierungshierarchien werden zur Transformation von Werten eingesetzt

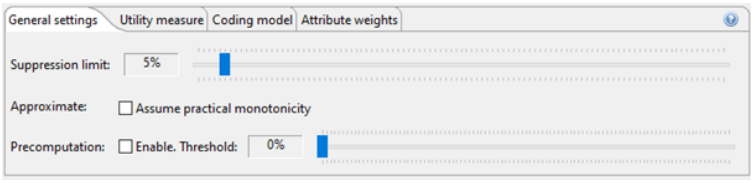


- In ARX werden Generalisierungshierarchien meist als Tabellen dargestellt
- Jede Zeile einer solchen Tabelle enthält die Abbildungsregel für einen Attributwert
- Für die hier gezeigte Hierarchie ergibt sich bspw. folgende Darstellung:

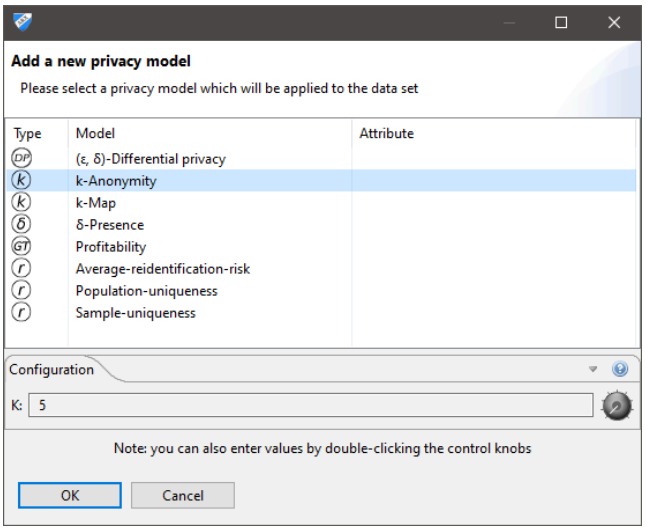
Stufe 0	Stufe 1	Stufe 2
1	1-29	*
...	...	...
29	1-29	*
30	30-59	*
...	...	...

- **Hinweis:** Für die sichere Durchführung von Anonymisierungsverfahren mit ARX ist es zwingend notwendig, dass jedes in einem Datensatz abgebildete Individuum durch genau einen Eintrag repräsentiert ist!

3. Schützen Sie nun die Einträge des Datensatzes vor Re-Identifikation. Spezifizieren Sie ein maximales Re-Identifikationsrisiko von 20% für jeden Eintrag im Datensatz. Verwenden Sie Generalisierung von Attributen und Unterdrückung von Einträgen zum Schutz des Datensatzes. Stellen Sie sicher, dass nicht mehr als 5% der Einträge entfernt werden.
   Wählen Sie unter “General Settings” ein Unterdrückungs Limit von 5 %. Das ist die Begrenzung des Anteils von Einträgen die entfernt werden können.



   Spezifizieren Sie nun die Risikogrenzwerte / Datenschutzmodelle. Fügen Sie unter der Liste aller aktiver Schutzmodelle ein neues Schutzmodell hinzu. Wählen Sie k-Anonymity mit k=5 (Risikogrenzwert 1/5=20%).



4. Anonymisieren Sie nun den Datensatz über die Menüleiste (✅). Nach erfolgreicher Anonymisierung werden in der rechten oberen Ecke von ARX Informationen über den Prozess angezeigt

Hintergrund

- Eine Transformation der Eingabedaten verringert die Eindeutigkeit von Wertekombinationen
- Dadurch entstehen Gruppen von Einträgen, auch Klassen genannt

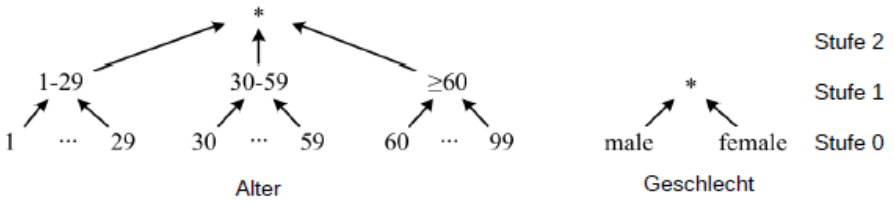
- Die Größe der Gruppe in der sich der zu einem Individuum gehörende Datensatz befindet bestimmt die Wahrscheinlichkeit mit der der Eintrag dem Individuum zugeordnet werden kann
- Befindet sich ein Eintrag beispielsweise in einer Gruppe der Größe k, so kann dieser nur mit einer Wahrscheinlichkeit von 1/k dem Individuum richtig zugeordnet werden
- Man nennt diese Wahrscheinlichkeit das Re-Identifikationsrisiko
- Die Attribute die zur Gruppenbildung herangezogen werden nennt man Quasi-Identifikatoren
- Das Datenschutzmodell k-Anonymity fordert für jeden Eintrag eine Klassengröße von mindestens k
- Das bedeutet, dass k-Anonymity das Re-Identifikationsrisiko jedes Eintrags auf maximal 1/k beschränkt
- Die Transformation von Eingabedaten erfolgt meist durch das Generalisieren von Attributwerten und das Entfernen (auch "Unterdrücken" genannt) von Einträgen
- Üblicherweise wird eine Obergrenze für die Anzahl der Einträge die entfernt werden dürfen festgelegt („Suppression limit“)

Aufgabe 2)

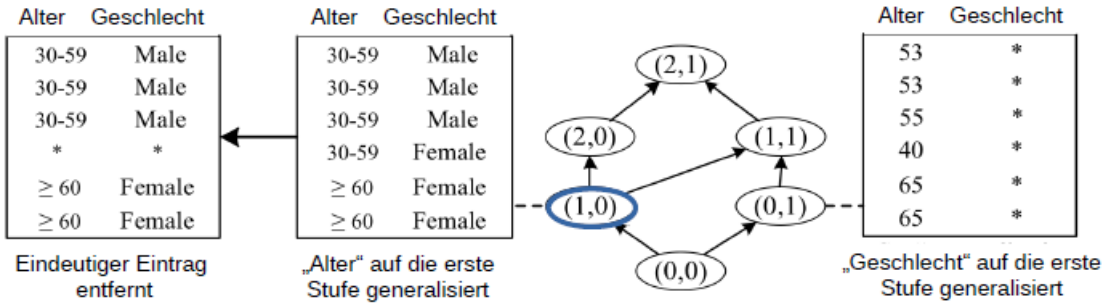
Untersuchen Sie den Lösungsraum. Finden Sie alternative Lösungen mit einem nahezu optimalen Informationsverlust. Untersuchen Sie Unterschiede zur vorhergehenden Lösung.

Hintergrund

- Generalisierungshierarchien

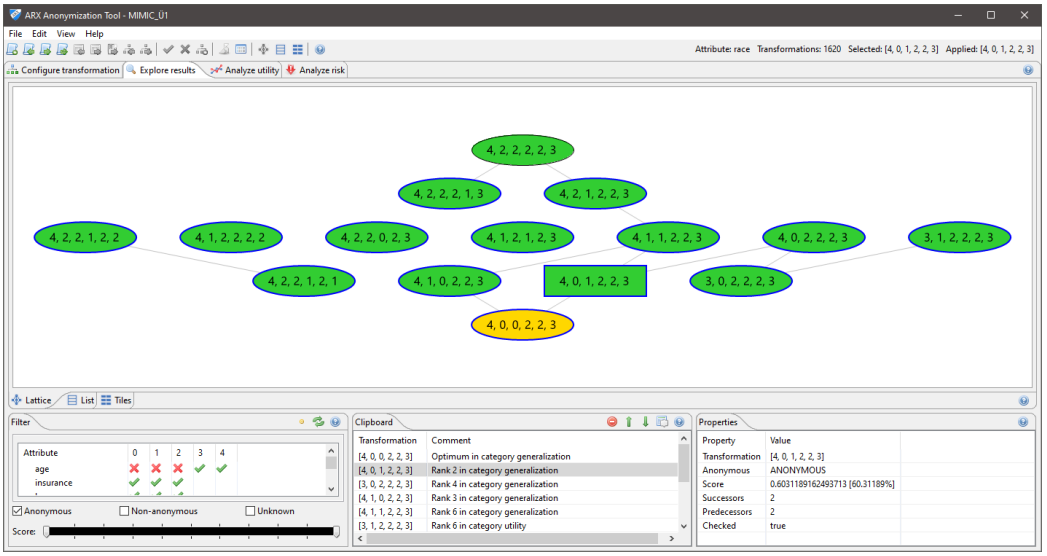


- Lösungsraum

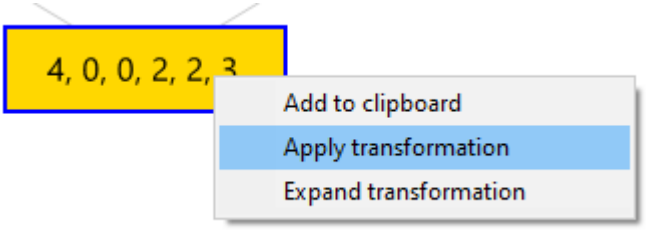


- Jede Transformation (im Bild markiert) im Lösungsraum definiert eine eindeutige Kombination von Generalisierungsstufen, eine für jeden Quasi-Identifikator
- Zusätzlich werden oftmals Einträge automatisch entfernt („suppression")

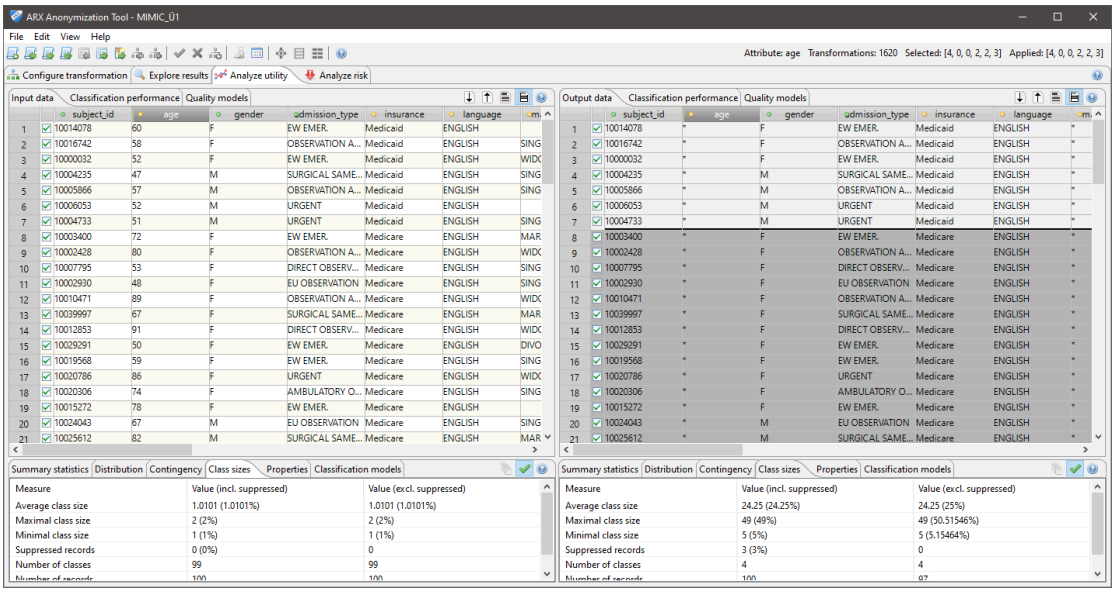
1. Verwenden Sie die Explorationsperspektive von ARX:
   - Grün: anonyme Transformation
   - Gelb: optimale Lösung
   - Rechteck: ausgewählte Transformation



Hinweis: Mit einem Rechtsklick auf eine Transformation können Sie „Apply transformation“ wählen, um eine ausgewählte Transformation auf den Datensatz anzuwenden



2. Wechseln Sie nun in die Qualitätsanalyseperspektive von ARX:
   - Links: Eingabe Daten
   - Rechts: Ausgabe der angewandten Transformation



3. Schauen Sie sich den Reiter „Class sizes“ an und beantworten Sie die folgenden Fragen:

Füllen Sie aus:

- Optimale Transformation
  - [_, _, _, _, _, _, _, _, _]
  - Eigenschaften
    - Vollständig entfernte(s) Attribut(e): ___ (_____)
    - Unterdrückte Einträge: ____ (__,_%)
    - Informationsverlust: __,_%
    - Durchschn. Klassengröße / Average class size: __,_
- Ihre Auswahl
  - [_, _, _, _, _, _, _, _, _]
  - Eigenschaften
    - Vollständig entfernte(s) Attribut(e): ___ (_____)
    - Unterdrückte Einträge: ____ (__,_%)
    - Informationsverlust: __,_%
    - Durchschn. Klassengröße / Average class size: __,_

Aufgabe 3)

Analysieren Sie Re-Identifikationsrisiken von Einträgen in einem 5-anonymen Datensatz. Hier entspricht 20% 5-anonymity. Untersuchen Sie wie sich Re-Identifikationsrisiken durch die Anonymisierung verändert haben.

Hintergrund:

- Die Größe der Gruppe in der sich ein Eintrag befindet bestimmt die Wahrscheinlichkeit mit der er dem zugehörigen Individuum richtig zugeordnet werden kann, d.h. das Re-Identifikationsrisiko
- Selbst bei einem k-anonymen Datensatz muss davon ausgegangen werden, dass durch zufälliges Abbilden von Individuen in passende Gruppen ein nicht geringer Anteil der Einträge richtig zugeordnet werden kann
- Diese Betrachtungen setzen voraus, dass der Angreifer bereits weiß, dass Daten über die Individuen die er versucht abzubilden im Datensatz vorhanden sind
- Man nennt dieses Worst-Case Szenario das Prosecutor Modell

1. Wechseln Sie in den Reiter „Analyze Risk“ und wählen Sie “Attacker models”:

Füllen Sie aus:

– Prosecutor Re-Identifikationsrisiko für den Eingabedatensatz

- __,_% der Einträge sind einem Risiko von mehr als 20% ausgesetzt
- Die zu erwartende relative Anzahl an korrekt re-identifizierten Einträgen ist __,_%
- Das niedrigste Risiko ist ___%, das höchste Risiko ist ___%
- Das höchste Risiko betrifft ___% der Datensätze

– Prosecutor Re-Identifikationsrisiko für den Ausgabedatensatz

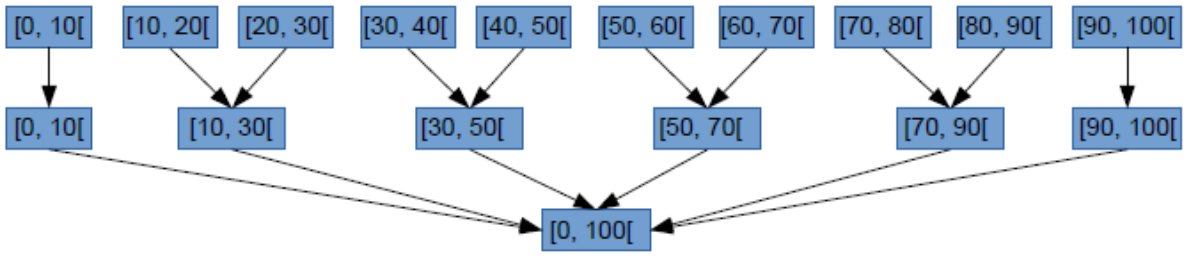
- __,_% der Einträge sind einem Risiko von mehr als 20% ausgesetzt
- Die zu erwartende relative Anzahl an korrekt re-identifizierten Einträgen ist __,_%
- Das niedrigste Risiko ist ___%, das höchste Risiko ist ___%
- Das höchste Risiko betrifft ___% der Datensätze

Feststellung 1: _____

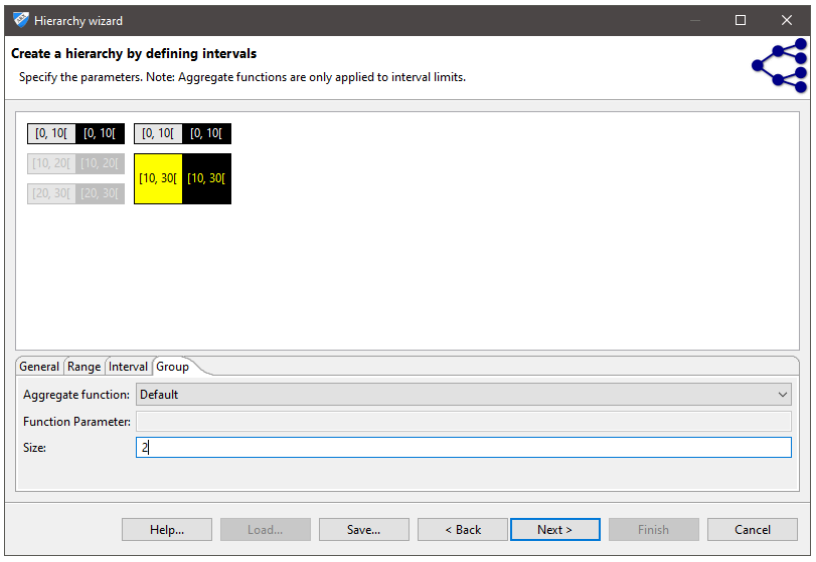
Feststellung 2: _____

Aufgabe 4)

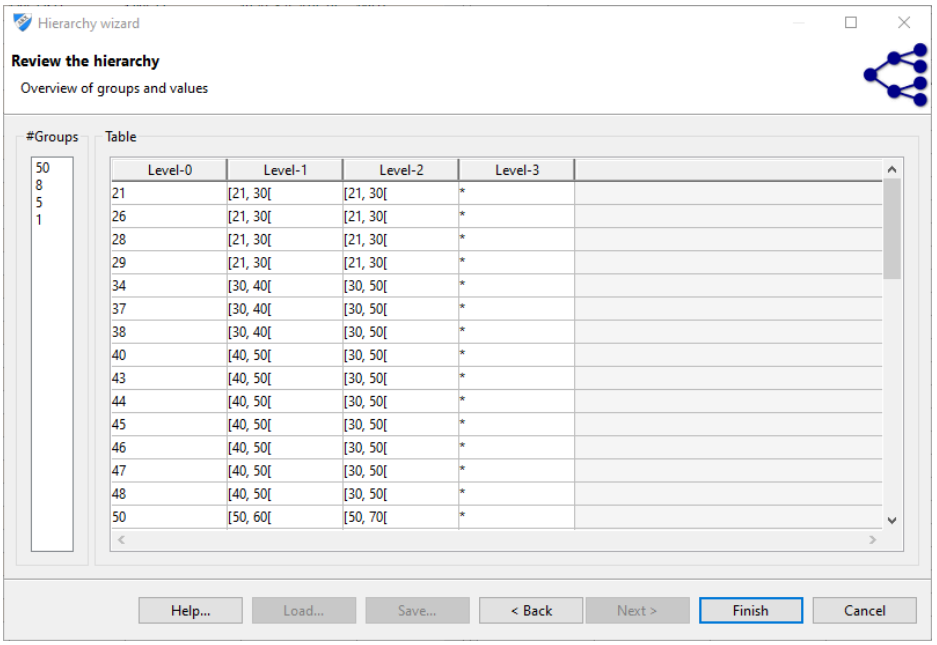
Erstellen Sie eine Generalisierungshierarchie für das Attribut „age“ die folgende Struktur besitzt:



1. Wählen Sie „Edit → Create hierarchy...“ aus dem Anwendungsmenü.
   Mit ARX können Sie drei Arten von Hierarchien erzeugen:
   - Basierend auf Intervallen
     Beispiel: Alter, Laborwerte
   - Basierend auf einer Ordnung der Attributwerte
     Beispiel: Kodierte Diagnosen
   - Durch Maskierung
     Beispiel: Postleitzahl
2. Wählen Sie „Use intervals“ und klicken Sie „Next“. Wählen Sie nun den Reiter "Range":
   Setzen Sie die untere Schranke auf "0" und die obere Schranke auf "100".
3. Klicken Sie anschließend auf das eine angezeigte Intervall. Definieren Sie ein Basisintervall der Größe 10:
   Setzen Sie hier die untere Schranke auf "0" und die obere Schranke auf "10".
4. Fügen Sie nun eine weitere Generalisierungsstufe ein.
   4.1. Machen Sie einen Rechtsklick auf das Basisinterval und wählen Sie „Add new level“.
   4.2. Dann wählen Sie mit einem Rechtklick auf den neu entstandenen Knoten „Add after“ um einen Geschwisterknoten auf der gleichen Ebene einzufügen.
   4.3. Um das Interval]10, 30] zu definieren, müssen zwei Intervalle der darunterliegenden Ebene zusammengefasst werden. Definieren Sie demnach eine Größe von 2:



5. Erzeugen Sie anschließend die Intervalle [30, 50[, [50,70[, [70, 90[, [90, 100[.



Aufgabe 5)

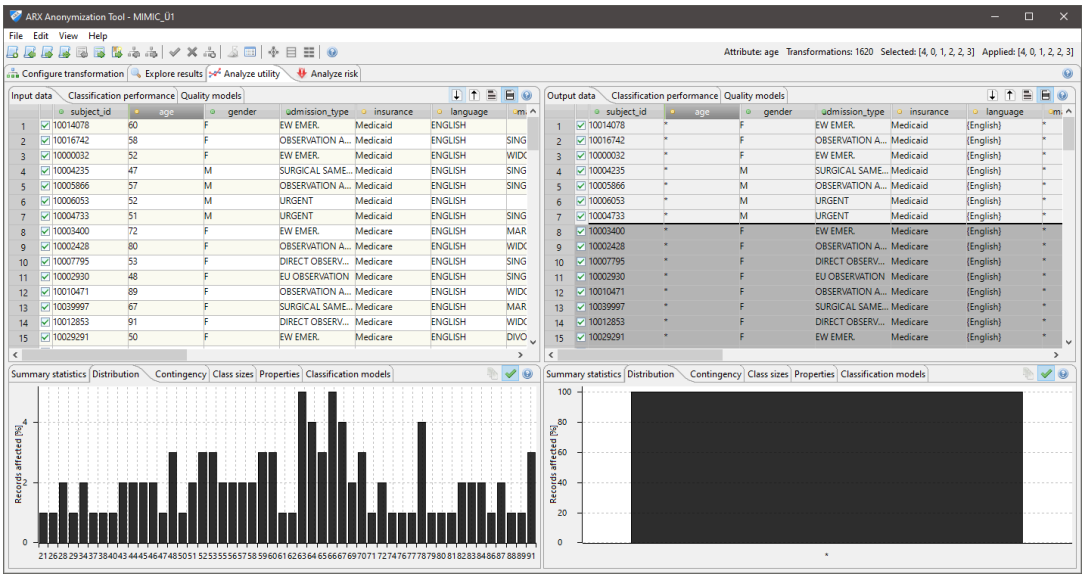
Machen Sie sich weiter mit ARX vertraut. Untersuchen Sie…

- ... die Verteilung des Attributs „age“
- ... statistische Parameter der Verteilung des Attributs „age“
- ... die Kontingenz zwischen „age“ und „marital-status“
- ... die Auswirkung des Anonymisierungsprozesses auf die Vorhersagbarkeit des Attributs „marital-status“

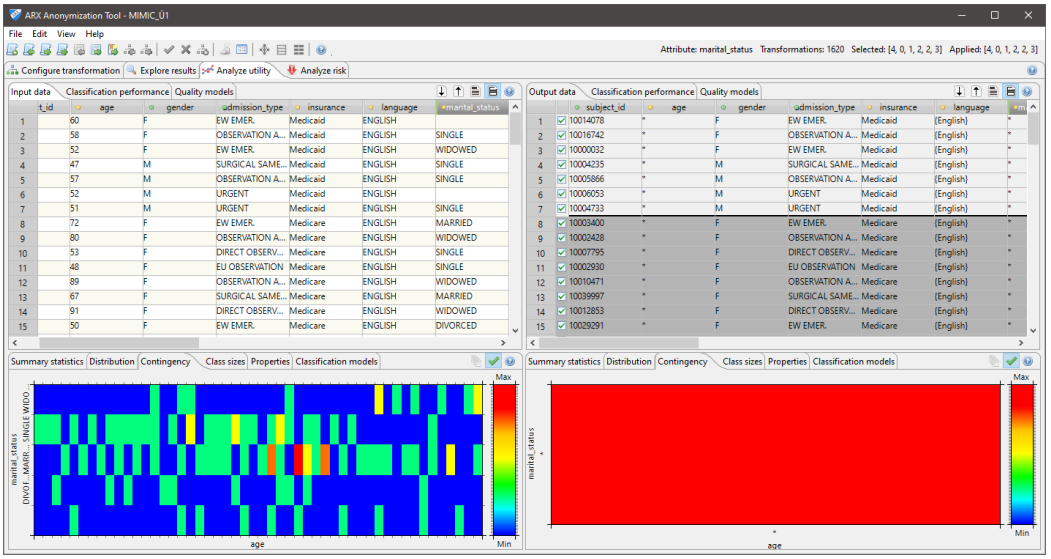
Hintergrund:

- Da es oft nur schwer möglich ist, automatisiert eine angemessene Lösung für ein Anonymisierungsproblem zu finden, werden von ARX verschiedene Funktionalitäten zur Analyse der Nützlichkeit der Ausgabedaten bereitgestellt
- Zur Untersuchung der Auswirkungen des Anonymisierungsprozesses auf ein häufiges Anwendungsszenario implementiert ARX eine Methode aus dem Bereich maschinelles Lernen
- Hierzu wird ein logistisches Regressionsmodell so trainiert, dass es auf Basis einer vorgegebenen Menge von Feature-Attributen den Wert eines Klassenattributs für Einträge im Eingabedatensatz vorhersagen kann
- Die Auswirkung des Anonymisierungsprozesses zeigen sich in einem Rückgang der Genauigkeit des Klassifikators wenn er statt mit den Eingabedaten, mit den anonymisierten Ausgabedaten trainiert wurde
- Die von ARX angezeigten Werte sind als Richtwerte zu sehen

1. Wechseln Sie in die Qualitätsanalyseperspektive von ARX. Selektieren Sie das Attribut „age“ durch einen Mausklick auf die enstprechende Spalte und dann den Reiter „Distribution“.

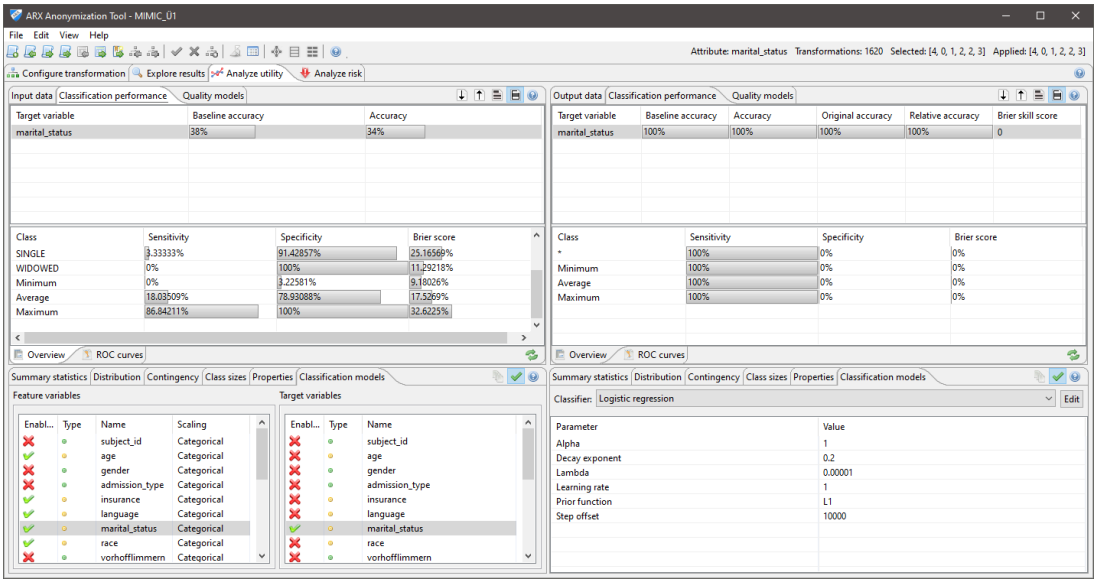


2. Selektieren Sie den Reiter „Summary statistics“ und vergleichen Sie die “Scale of measure”. Selektieren Sie erst das Attribut „age“, dann das Attribut „marital-status“ und wählen Sie dann den Reiter „Contingency“.



Die Kontingenztabelle ist als Heatmap dargestellt. Vergröberung durch Generalisierung von Attributwerten. Verzerrung durch Unterdrückung von Einträgen. Alle Einträge von martial_status wurden entfernt.

3. Wählen Sie nun den Reiter „Classification performance“. Der Klassifikator sagt Werte der Class-Attribute auf Basis der Feature-Attribute vorher. Vergleichen Sie nun die Genauigkeit des Klassifikators vor & nach dem Anonymisierungsprozess.



Hier ist die Accuracy (Relative Genauigkeit) ist von 35 % auf 100% gestiegen.

Aufgabe 6)

Analysieren Sie die Auswirkung von Unterdrückung („Suppression“) auf den Informationsverlust. Anonymisieren Sie den Datensatz mit unterschiedlichen Grenzwerten für die Anzahl der maximal zu entfernenden Einträge („Suppression limit“).

Hintergrund:

- Generalisierung von Attributwerten reduziert die Eindeutigkeit von Einträgen
- Häufig gibt es Einträge in einem Datensatz, die nur durch die Verwendung einer sehr hohen Generalisierungsstufe an Eindeutigkeit verlieren (Ausreisser)
- Um zu verhindern, dass der gesamte Datensatz stark generalisiert werden muss, können diese Einträge automatisch entfernt werden
- Dadurch kann der Verlust an Information reduziert werden
- **Hinweis**: Hier zeigt sich eine grundlegende Eigenschaft vieler Anonymisierungsverfahren. Meist werden häufig auftretende „Muster“ extrahiert.

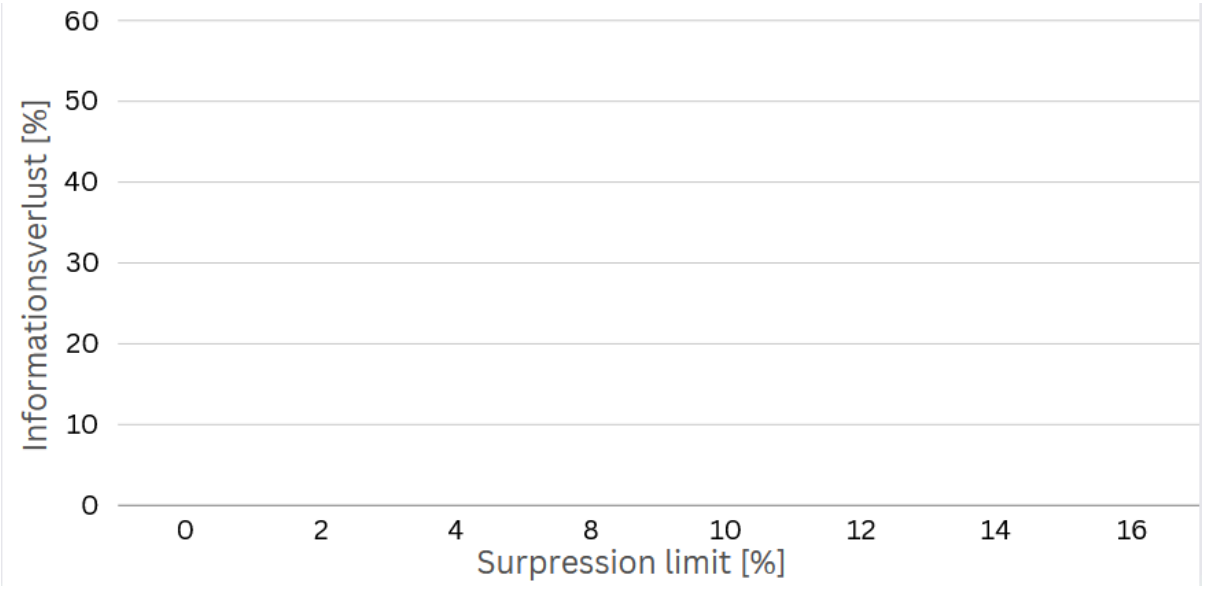
1. Wechseln Sie in die Konfigurationsperspektive von ARX
   - Bestimmen Sie den maximalen Grenzwert der sich zu untersuchen lohnt. Anonymisieren Sie den Datensatz dazu mit einem "suppression limit“ von 100%

   Wechseln Sie in die Qualitätsanalyseperspektive von ARX
   - Der Grenzwert entspricht der relativen Anzahl der unterdrückten Einträge (aufgerundet). Sie finden diese Kenntzahl im Reiter „Class sizes“.

2. Anonymisieren sie den Datensatz anschließend mit „suppression limits“ von 0%, 2%, ..., bis einschließlich dem bestimmten Grenzwert (aufgerundet).

Füllen Sie aus:

- Der Grenzwert ist ___%



Feststellung 1: _____

Feststellung 2: _____

Aufgabe 7)

Analysieren Sie die Auswirkung von Mikroaggregation auf die Ausgabedaten. Vergleichen Sie dazu Mikroaggregation und Generalisierung. Betrachten Sie das Attribut „age“.

Hintergrund:

- Generalisierung und Mikroaggregation verringern die Eindeutigkeit von Werten
- Bei Generalisierung wird auf einen gemeinsamen allgemeineren Wert abgebildet:
  50 → [50,70]
  60 → [50,70]
  70 → [50,70]
- Mikroaggreation ist ein Verfahren für numerische Attribute, bei dem Werte durch ein Aggregat ersetzt werden:
  50 → Avg(50,60,70) = 60
  60 → Avg(50,60,70) = 60
  70 → Avg(50,60,70) = 60

1. Anonymisieren Sie den Datensatz mit verschiedenen Transformationsmethoden für "age"
   1. Generalisierung
   2. Mikroaggregation mit Modus („Mode“)

   Wechseln Sie nach jedem Anonymisierungsvorgang in die Qualitätsanalyseperspektive. Notieren Sie die Paramter aus den Reitern „Summary statistics“ und „Class sizes“.

Füllen Sie aus:

	Input	Generalization	Microaggregation
Scale of measure			
Mode			
Median			
Min			
Max			
Suppressed records			

Feststellung 1: _____

Feststellung 2: _____