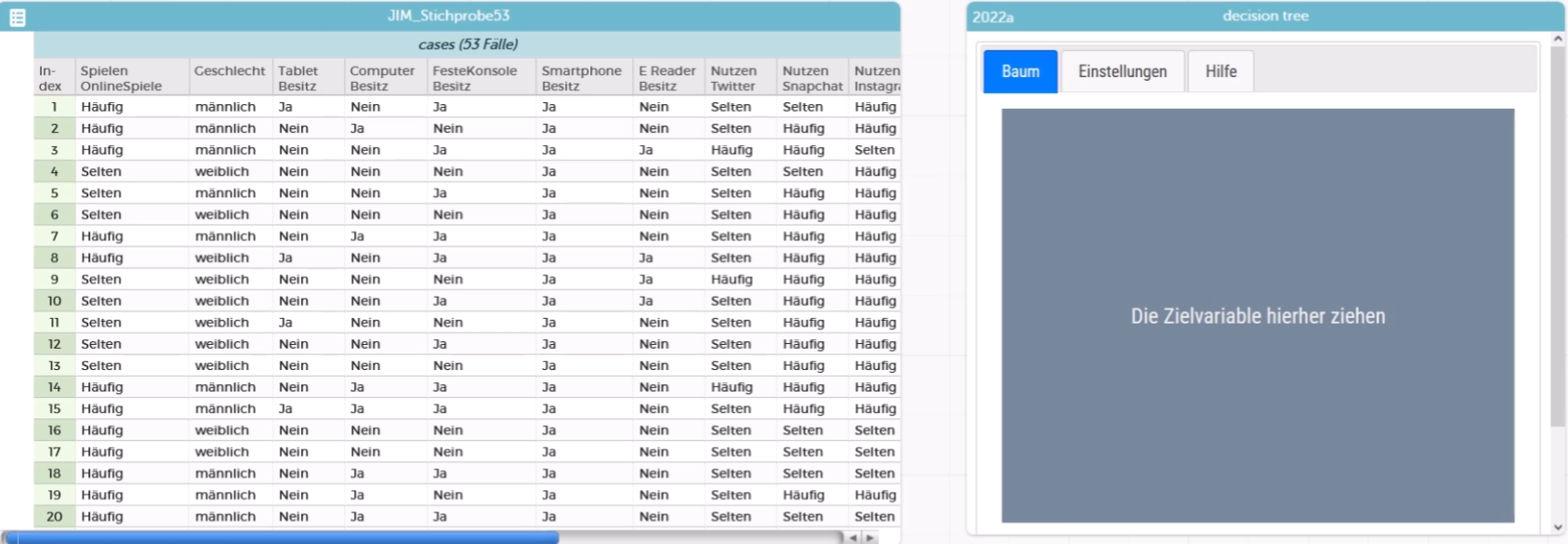
Einen Entscheidungsbaum mit dem Plug-In Arbor in CODAP erstellen

**Link zu CODAP:** https://tinyurl.com/CODAPEntscheidungsbaum

**1. Zielvariable festlegen**

Im Folgenden wird gezeigt wie in CODAP ein Entscheidungsbaum erstellt werden kann, der Personen anhand ihres Medienverhaltens klassifiziert. Der Baum soll vorhersagen, ob eine Person **Häufig** Onlinespiele spielt oder nur **Selten**. Um einen solchen Entscheidungsbaum zu erstellen müssen wir in CODAP zunächst die Zielvariable **Spielen\_Onlinespiel** auswählen, indem wir sie per Drag & Drop in das Entscheidungsbaum Plug-In ziehen.



**Drag & Drop**

Das nächste Bild zeigt das Entscheidungsbaum Plug-In mit ausgewählter Zievariable. Die Zielvariable sowie der Positiv-Wert werden oben angezeigt.

* Die **Zielvariable** (Spielen\_OnlineSpiele) ist diejenige für die eine Ausprägung vorhergesagt werden soll. (Ausprägungen: Häufig/Selten)
* Der **Positiv-Wert** ist hier die Ausprägung Häufig und er dient dazu die weiteren angezeigten Werte „32 zu 21“ interpretieren zu können. Im Datensatz sind insgesamt 53 Fällen, von denen 32 Häufig Onlinespiele spielen und 21 nur Selten. Die erste angezeigt Anzahl bezieht sich also immer auf den Positiv-Wert.



**Positiv - Wert**

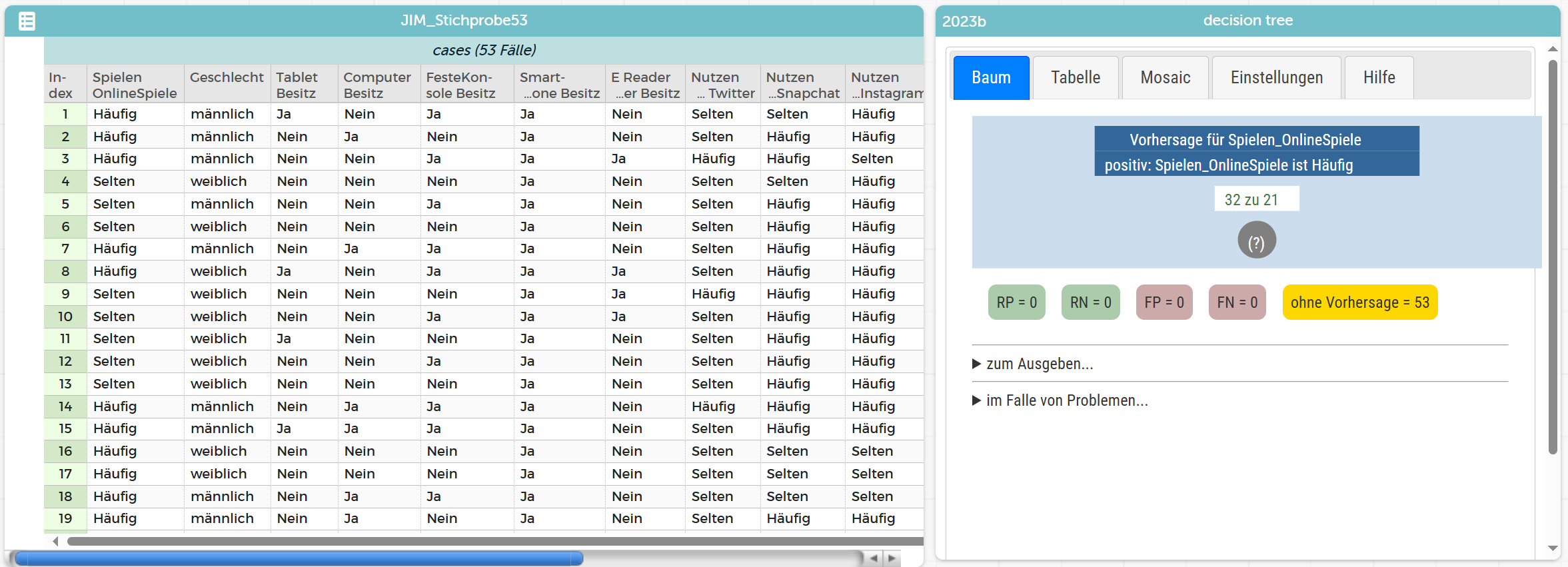
**Zielvariable**

Die Zielvariable kann wieder geändert werden, indem man eine andere Variable aus dem Datensatz in das obere dunkelblaue Rechteckt zieht, in dem die aktuelle Zielvariable angezeigt wird.

**2. Einstufigen Entscheidungsbaum erstellen**

Nun können wir die weiteren Merkmale, die im Datensatz angezeigt werden, als **Prädiktorvariablen** nutzen, um Entscheidungsregeln zu definieren. Dafür können die Variablen per Drag & Drop in das helle Rechteck unterhalb der Zielvariable eingefügt werden.

**Prädiktor-variablen**



**Drag & Drop**

ss

Eine Entscheidungsregel wird automatisch auf Grundlage der Ausprägungen der ausgewählten Variable erstellt. Das nennt man Datensplit. Hier wird der Datensatz Datensplit anhand des Geschlechts in 26 Konsolenbesitzer und 27 Nicht-Konsolenbesitzer aufgeteilt. Für diese Untergruppen werden die **Häufigkeiten** bezüglich der Zielvariable angezeigt:

* Von den Konsolenbesitzern spielen 20 **Häufig** Onlinespiele und 6 **Selten**.
* Von den Nicht-Konsolenbesitzern spielen 12 **Häufig** Onlinespiele und 15 **Selten**.

**Positiv - Wert**



**Häufigkeiten**

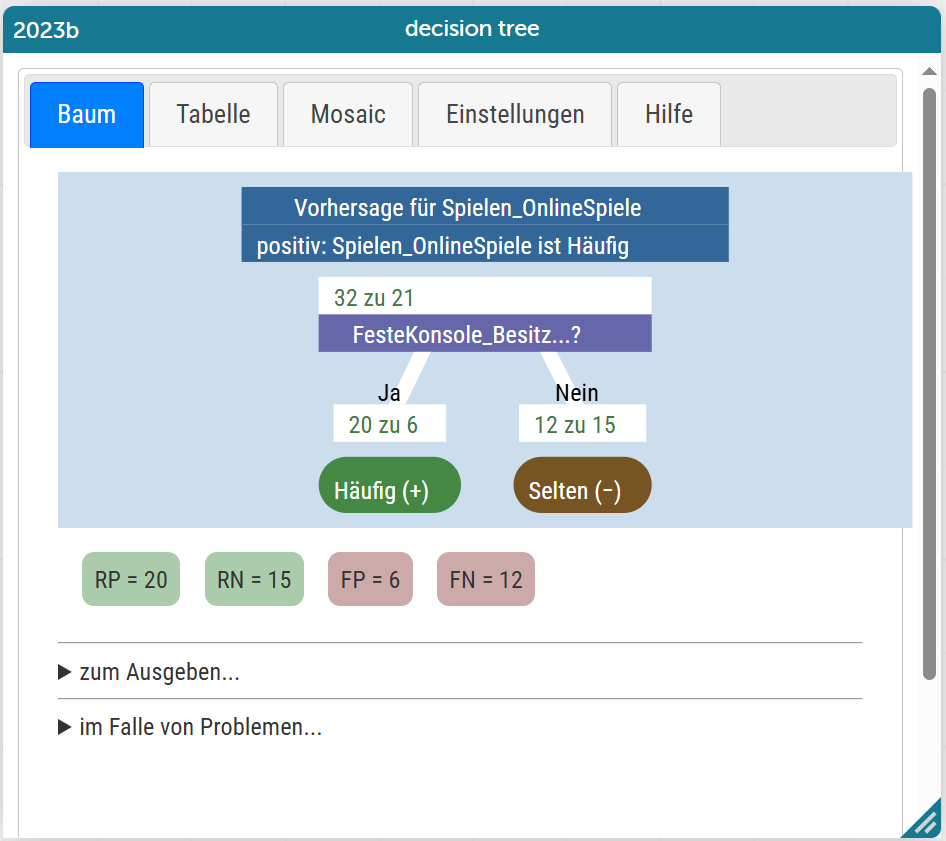
**wiederholt anklicken**

**Tipp:** Wenn man sich nicht mehr sicher ist ob z. B. die 20 zum Wert **Häufig** oder **Selten** gehört, kann man die Maus einfach etwas über dem Rechteck schweben lassen und man bekommt dort einen Info-Text angezeigt.

Um für „Blattknoten“ am Ende eines Pfades einen Entscheidungswert (Label) festzulegen, können wir die Blattknoten durch wiederholtes anklicken mit einem Entscheidungswert (Label) versehen. Dabei können wir uns an den angezeigten Werten orientieren. Im Bild sehen wir, dass 20 von 26 Fällen im linken Pfad den Wert **Häufig** enthalten. Aufgrund dieser großen Mehrheit wählen wir **Häufig** als Entscheidungswert für den Blattknoten.

**3. Entscheidungsbaum evaluieren**

Wenn wir einen Entscheidungsbaum erstellt und Werte für die Blattknoten festgelegt haben, können wir unten in der Trefferübersicht kontrollieren, wie gut der Baum für die Trainingsdaten funktioniert.



**Treffer-übersicht**

Die Abkürzungen RP, RN, FP, FN bedeuten:

* RP – Richtig Positive
* RN – Richtig Negative
* FP – Falsch Positive
* FN – Falsch Negative

Der Positiv-Wert ist hier „Häufig“ und Richtig Positive Fälle sind Fälle für die die Zielvariable die Ausprägung „Häufig“ hat und auch durch den Baum auch so klassifiziert wird. Hier sind das 20 Fälle. Falsch Positive sind Fälle die als „Häufig“ vorhergesagt werden, für die aber eigentlich „Selten“ richtig ist. Hier sind das 6 Fälle. Für Richtig Negative und Falsch Negative ist das analog nur mit der Ausprägung „Selten“.

Die Trefferübersicht kann unter dem Reiter „Tabelle“ als etwas detailliertere so genannte   
Konfusionsmatrix angezeigt werden. Dies ist eine Kreuztabelle, wo vorhergesagter Wert durch die Zeile angezeigt wird und der tatsächliche Wert durch die Spalte. In den entstehenden Zellen stehen die Häufigkeiten der jeweiligen Kombination. Somit werden dort die vier Werte RP, RN, FP, FN dargestellt:

