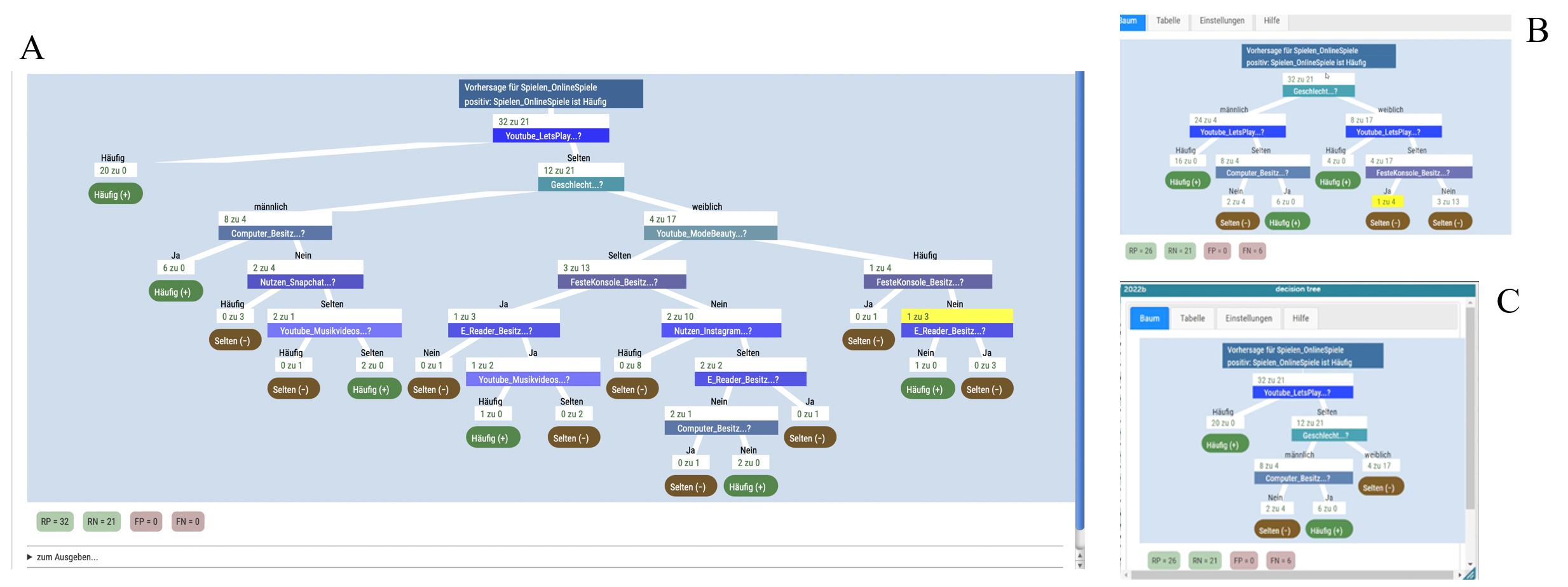
1. **Hintergrund**

Die folgenden Ausführungen basieren auf einer Studie, die mit drei Kohorten von Lernenden durchgeführt wurde. Insgesamt wurden 25 Paare von Lernenden untersucht bei der Bearbeitung von *Arbeitsblatt\_3\_Entscheidungsbaum* und der anschließenden Präsentation der erstellten Entscheidungsbäume, sowie der gewählten Ansätze für die Vorgehensweise. Wir wollten damit erfassen, welche Vorgehensweisen bei dieser sehr freien Aufgabe vorkommen, um daraus abzuleiten, wie man damit weiter umgehen kann.

1. **Eigenschaften der erstellten Entscheidungsbäume**

Die Entscheidungsbäume der Schüler waren sehr vielfältig. Damit ein Entscheidungsbaum eine angemessene Lösung der Aufgabe ist, muss er mindestens eine Prädiktorvariable enthalten und vorhergesagte Labels an jedem Endpunkt der Äste enthalten. Einer der 25 Entscheidungsbäume stellte keine angemessene Lösung dar, da keine vorhergesagten Labels spezifiziert wurden. Die übrigen 24 Entscheidungsbäume waren angemessen und haben zwischen 2 und 12 Prädiktorvariablen verwendet, was zu Anzahlen von Fehlklassifikationen zwischen 0 und 8 führte. Wie in Abbildung 4 dargestellt, erstellten die Lernenden sehr unterschiedliche Bäume, sodass von der Gestalt her sehr große Entscheidungsbäume (A), mittelgroße Entscheidungsbäume (B) oder kleine Entscheidungsbäume (C) entstanden.

Abbildung 1: Beispiele für Entscheidungsbäume von Schülern in CODAP

1. **Ansätze zur Auswahl von Prädiktorvariablen**

Bei der Analyse der mündlichen Präsentationen der Lernenden identifizierten wir zwei übergeordnete Ansätze zur Auswahl von Prädiktorvariablen:

* **Datenbasierte Argumentation** (16 Paare), bei der sich die Lernenden entweder auf Häufigkeitsverteilungen beziehen, die in der Entscheidungsbaum Darstellung angezeigt werden, oder auf Beobachtungen die sie direkt in der Datentabelle machen.
* **Kontextbasierte Argumentation** (8 Paare), bei der die Schüler ihr Kontextwissen nutzten, um Variablen zu identifizieren, die potenziell mit dem Spielen von OnlineGames zusammenhängen könnten.

Die kontextbezogene Argumentation basiert auf Überlegungen, die nicht von den Daten abhängen. Ein Beispiel für diesen Ansatz:

**Aussage:** „Ich habe mit den Geräten angefangen. Also zuerst, ob ein Computer vorhanden ist, dann, ob eine Konsole vorhanden ist, dann ein Tablet und dann ein Smartphone. Einfach, weil ich dachte, dass das wahrscheinlich am meisten Sinn ergibt: Wenn jemand ein Gerät hat, dann spielt er vermutlich auch darauf.“

Diese Argumentation beruht auf kontextbezogenen Annahmen über die Relevanz des Gerätebesitzes für die Vorhersage der OnlineGames-Nutzung. Es wurde nicht auf die vorhandenen Daten verwiesen.

Ein Beispiel für datenbasierte Argumentation ist der folgende Ausschnitt:

**Aussage:** „Mein Ziel war eigentlich immer, äh, ich habe gar nicht wirklich über die Namen der Variablen nachgedacht, sondern bin einfach alle durchgegangen und habe geschaut, wie ich ein gutes Verhältnis auf beiden Seiten erreichen kann.“

In diesem Beispiel wurden verschiedene Datensplits mit unterschiedlichen Prädiktorvariablen analysiert und die resultierenden Häufigkeitsverteilungen betrachtet, ohne die Bedeutung der Variablen zu berücksichtigen – ähnlich wie es professionelle Entscheidungsbaum-Algorithmen tun. Wir haben Unterkategorien der datenbasierten Auswahl von Prädiktorvariablen identifiziert:

* **Schrittweise Reduktion der Fehlklassifikationen** (9 Paare), wobei das Ziel darin besteht, die Anzahl der Fehlklassifikationen in jedem Schritt so stark wie möglich zu reduzieren.
* **Schrittweise Identifikation von „n zu 0“** (5 Paare), bei der nach einem Datensplit gesucht wird, die mindestens eine reine Teilmenge ergibt (Häufigkeitsverteilung „n zu 0“ oder „0 zu n“), unabhängig von der Anzahl der Fehlklassifikation.
* **Aufmerksame Beobachtung der Daten** (2 Paare), bei der die Lernenden die Datentabelle genau untersuchten, um Zusammenhänge zwischen Prädiktoren und der Zielvariablen zu erkennen.

Einige Paare haben kontextbasierte und datenbasierte Argumentation kombiniert, indem sie zunächst mit Kontextwissen Variablen vorausgewählt und sie anschließend datenbasiert angeordnet haben; diese Mischformen haben wir als (dominant) kontextbasierten Ansatz kategorisiert.

1. **Ansätze für Abbruchkriterien**

Wir haben vier verschiedene Typen von Abbruchkriterien bei den Lernenden identifiziert:

* **Erreichen von null Fehlklassifikationen**
* **Stagnation der Anzahl der Fehlklassifikationen**
* **Anzahl der Fälle in einer Teilmenge „zu stark“ reduziert**
* **kontextbasiertes Abbruchkriterium**

Ein häufiges datenbasiertes Kriterium war das **Erreichen von null Fehlklassifikationen** (12 Paare), wie in diesem Beispiel:

**Aussage:** „Das Stoppkriterium war in diesem Fall tatsächlich, äh, wenn in jedem Blatt ein 0-zu-n- oder n-zu-0-Verhältnis besteht, dann höre ich auf, weil ich keine Fehler mehr habe.“

Hier beendet der Schüler den Prozess, sobald keine Fehlklassifikationen mehr auftreten.

Andere Schüler nutzten die **Stagnation der Anzahl der Fehlklassifikationen** (7 Paare):

**Aussage:** „Dann, äh, stellte sich heraus, dass es für uns nicht sehr sinnvoll war, noch mehr hinzuzufügen, weil die Fehlerquote nicht mehr gesunken ist.“

In diesem Fall wurde aufgehört, sobald weitere Ergänzungen von Variablen zum Entscheidungsbaum nicht zu einer Reduktion der Fehlklassifikationen führte.

Ein dritter Typ begründete das Stoppen damit, dass eine weitere Datenaufteilung die **Anzahl der Fälle in einer Teilmenge** „zu stark“ reduzieren würde, um noch „repräsentativ“ zu sein (2 Paare):

**Aussage:** „Es waren noch 20 Personen darin enthalten, und deshalb war es noch relativ repräsentativ.“

Der Begriff „repräsentativ“ ist hier möglicherweise nicht ganz passend, aber es scheint, dass der Lernende an Generalisierbarkeit und die Klassifikation neuer Fälle jenseits der vorhandenen Daten dachte und so die Idee des Overfitting im Ansatz antizipiert.

Abschließend gibt es auch ein **kontextbasiertes Abbruchkriterium**, das sich auf die Verfügbarkeit von als relevant erachteten Variablen bezieht (3 Paare):

**Aussage:** „Wir wollten eigentlich nur das einbeziehen, was wirklich relevant und mit dem Spielen von Computerspielen verbunden war.“

Die Herangehensweisen an Abbruchkriterien zeigen unterschiedliche Prioritäten – sei es die Maximierung der Rate korrekt klassifizierter Fälle in den Trainingsdaten, das klassifizieren neuer Fälle außerhalb der Trainingsdaten oder die Sicherstellung der Kontextrelevanz.

1. **Typisierung der Zusammenhänge zwischen Prozessansätzen und Eigenschaften der erstellten Entscheidungsbäume**

Um Typen zu identifizieren, nutzten wir die Schüleransätze zur Auswahl von Prädiktorvariablen (Abschnitt 3) und Abbruchkriterien (Abschnitt 4), um die Eigenschaften ihrer Entscheidungsbäume (Abschnitt 2) zu erklären. Wir haben Untergruppen auf Basis der Kategorien aus 3. und 4. gebildet und wollten damit homogene Gruppen resultierender Entscheidungsbäume finden. Wir haben drei Typen identifiziert (Abbildung 2).

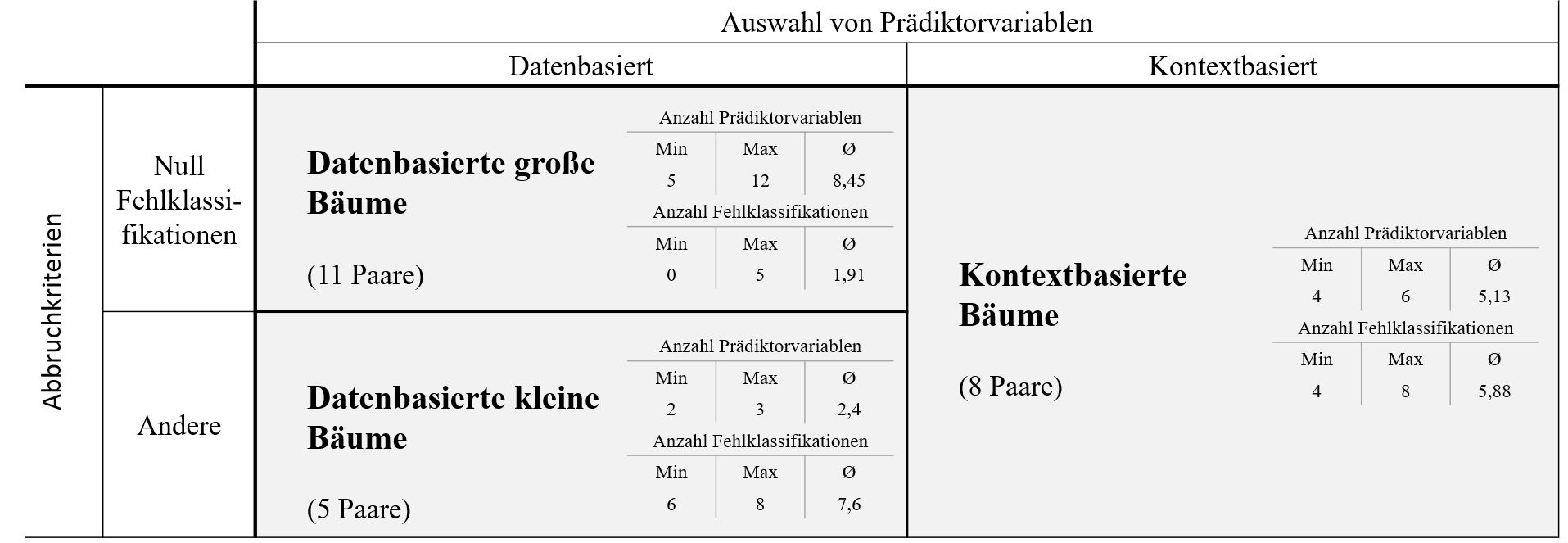


Abbildung 2: Typen von Ansätzen der Lernenden

Unsere Kategorien datenbasierte und kontextbasierte Auswahl der Prädiktorvariablen führten im Fall der **kontextbasierten Entscheidungsbäume** zu einer homogenen Untergruppe (4-6 Prädiktorvariablen, 4-8 Fehlklassifikationen).

Die datenbasierten Entscheidungsbäume (2-12 Prädiktoren, 0-8 Fehlklassifikationen) weisen hingegen eine hohe Variabilität auf, sodass diese weiter unterteilt werden. Eine hilfreiche Unterkategorie für datenbasierte Entscheidungsbäume ist das Abbruchkriterium – null Fehlklassifikationen vs. andere Kriterien – was zu zwei homogeneren Untergruppen führte:

* **Datenbasierte große Bäume**, die null Fehlklassifikationen anstreben   
  (5-12 Prädiktorvariablen, 0-5 Fehlklassifikationen).
* **Datenbasierte kleine Bäume** (2-3 Prädiktorvariablen, 6-8 Fehlklassifikationen).

Abbildung 1 zeigt Beispiele für einen **kontextbasierten Baum** (B), einen **datenbasierten großen Baum** (A) und einen **datenbasierten kleinen Baum** (C).

1. **Umgang mit den Unterschiedlichen Typen von Lösungen im Unterricht**

Wir identifizierten drei Typen von Ansätzen zur Erstellung von Entscheidungsbäumen. Abbildung 2 hilft Lehrkräften, diese Typen zu antizipieren und zu erkennen. Jeder Typ trat in allen drei untersuchten Kohorten auf, sodass plausibel ist, dass sie auch in weiteren Kohorten auftreten. Alle drei Typen bieten wertvolle Perspektiven für den Unterricht.

Tabelle 1: Häufigkeiten der Schüleransätze in den drei Kohorten

|  | | **Kontextbasierte Bäume** | **Datenbasierte große Bäume** | **Datenbasierte kleine Bäume** |
| --- | --- | --- | --- | --- |
| Kohorte 1 | 2 Paare | | 2 | 2 |
| Kohorte 2 | 2 Paare | | 5 | 1 |
| Kohorte 3 | 4 Paare | | 4 | 2 |

In der Klasse präsentieren alle Paare ihren Entscheidungsbaum, sowie den Ansatz für die Prädiktorvariablen Auswahl und Abbruchkriterien. Während die Lernenden ihre Ansätze präsentieren, kann die Lehrkraft schon Beispiele für die verschiedenen Typen notieren und anschließend die drei Typen von Lösungen zusammenfassen und an die gesehenen Beispiele knüpfen.

Im Folgenden geht es darum die Typen gegenüberzustellen, insbesondere **datenbasiertes Vorgehen vs. kontextbasiertes Vorgehen** und **große Bäume vs. kleine Bäume**. Man kann die Lernenden diskutieren lassen warum sie den ein oder anderen Typ bevorzugen. Hier sollte man erstmal alle Argumente zulassen und wenn alle Ansätze in der Klasse vertreten sind hat man meist auch für jeden Ansatz „Anwälte“.

Ein Argument für datenbasierte große Bäume kann sein, dass die beste Performance auf den Trainingsdaten vorliegt (starker Fokus auf Trainingsdaten), für datenbasierte kleine Bäume, dass nicht einfach wild lauter Variablen hinzugefügt wurden die keine gute Aussage für neue Fälle treffen (Fokus auf die Vorhersage neuer Fälle), und für kontextbasierte Bäume, dass nur Variablen benutzt wurden deren Zusammenhang mit der Zielvariable auch plausibel ist (Fokus auf Kontext). Zunächst wirkt es so, dass die datenbasierten großen Bäume die beste Lösung sind, da sie auch am meisten Arbeit benötigen. Später im Unterrichtsverlauf kann man einmal auflösen, dass die datenbasierten kleinen Bäume eigentlich die sinnvollste Lösung sind, da sie

Abschließend kann man zwei zentrale Fragen aufwerfen, die nicht sofort, aber in den folgenden Unterrichtsphasen beantwortet werden können:

1. **„Welcher Ansatz wird in professionellen Szenarien z. B. von Instragram genutzt – kontextbasiert oder datenbasiert?“**
2. **„Welcher Baumtyp eignet sich besser zur Klassifikation neuer Fälle – große oder kleine Bäume?“**

Die erste Frage führt zur Einführung datenbasierter ML-Algorithmen und die zweite zum Testen von Entscheidungsbäumen mit Testdaten und Overfitting. Somit sind die Themen der nächsten Unterrichtsphasen motiviert.