Aufgabenblatt: Entscheidungsbäume in Orange3

Im Folgenden wirst du in die Rolle eines Data Scientist schlüpfen, der mithilfe der zur Verfügung gestellten Daten der Äffchen aus dem Zoo ein KI-Modell zur Vorhersage von deren Beißverhalten trainiert. Ein Data Scientist arbeitet mit verschiedensten Daten und zieht mit Hilfe von wissenschaftlichen Analysen und entsprechender Programme Schlüsse aus diesen Daten. Das Modell, das wir heute lernen lassen, ist ein Entscheidungsbaum.

Öffne das Programm Orange3 auf deinem PC und lege ein neues Projekt an.

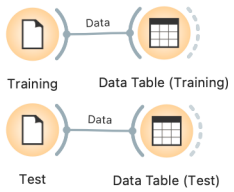
Aufgabe 1:

a) Von deiner Lehrkraft erhältst du zwei csv-Dateien, die die Trainings- (`affen-trainingsdaten.csv`) und Testdaten (`affen-testdaten.csv`) des Äffchenspiels beinhalten.

Öffne beide Dateien in Orange3, indem du das *File*-Widget aus dem Bereich *Data* auf die Leinwand ziehst.
Mit einem Doppelklick gelangst du in die Einstellungen. Dort kannst du die Datei auswählen (...). Schließe das Fenster wieder. Für jeden Datensatz benötigst du ein eigenes File-Widget: eins für die Trainingsdaten und eins für die Testdaten. Benenne die Widgets, indem du mit einem Rechtsklick das Kontextmenü öffnest und *Rename* auswählst (siehe auch Hilfskarte 1).

b) Noch kannst du die Daten aber nicht ansehen. Dafür wird ein weiteres Widget benötigt.

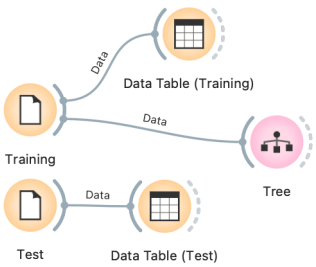
Ziehe das *Data Table*-Widget aus dem Bereich *Data* auf die Leinwand. Verbinde dann den Ausgang der *File*-Widgets mit einem *Data Table*-Widget. Mit einem Doppelklick auf das Widget kannst du dir die Daten ansehen (siehe auch Hilfskarte 2).
Beschreibe wie die Merkmale der Äffchen modelliert sind!



c) Jetzt hast du endlich die Daten, um einen **Entscheidungsbaum** lernen zu lassen.

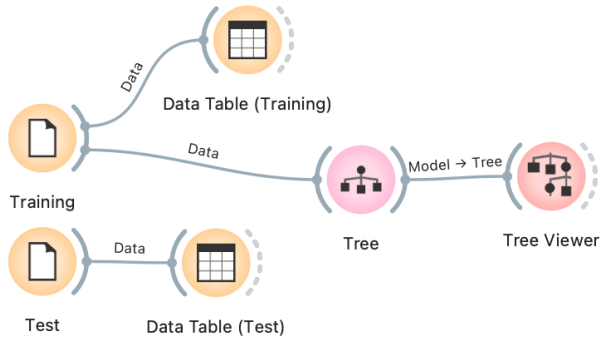
Ziehe dazu ein *Tree*-Widget aus dem Bereich *Model* auf die Leinwand. Verbinde die Trainingsdaten mit dem *Tree*-Widget.

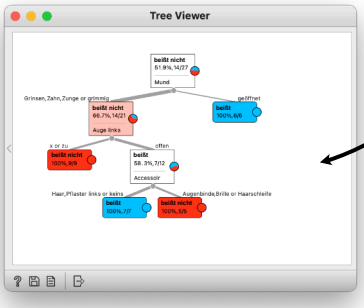
Klicke doppelt auf das Widget und stelle sicher, dass ein **binärer** Entscheidungsbaum gelernt wird und der Haken bei ☑ Induce binary tree gesetzt ist (siehe auch Hilfskarte 3).



d) Mit dem *Tree*-Widget kannst du den Baum aber noch nicht ansehen, sondern nur Einstellungen fürs Training festlegen.

Um den gelernten Baum anzuzeigen, ziehe ein *Tree Viewer*-Widget aus dem Bereich *Visualize* auf die Leinwand und verbinde es mit dem *Tree*-Widget. Mit einem Doppelklick kannst du dir den Baum ansehen (siehe auch Hilfskarte 4) .

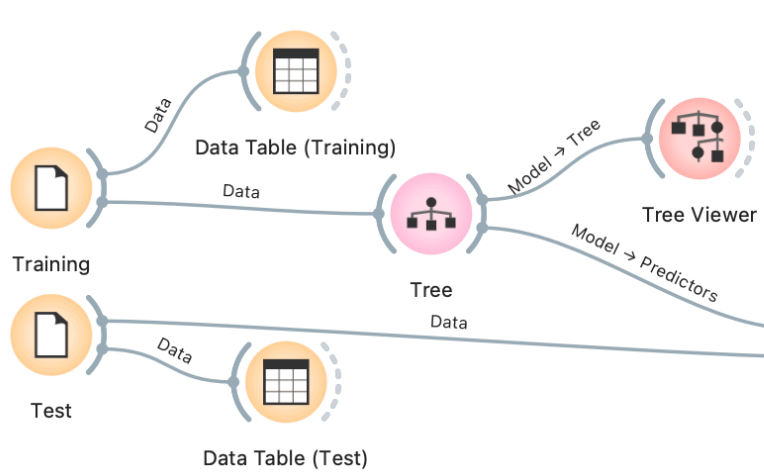


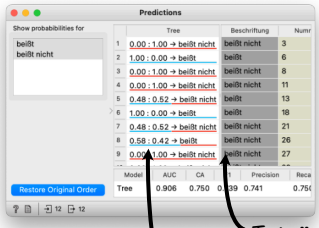


Dein Baum sollte ungefähr so aussehen.

e) Im Folgenden gilt es, den gelernten Entscheidungsbaum zu testen. Dazu stehen Testdaten bereit, auf die du nun den Baum anwendest.

Ziehe dazu ein *Predictions*-Widget aus dem Bereich *Evaluate* auf die Leinwand, das das zur Verfügung gestellte Modell auf die zu Verfügung gestellten Daten anwendet. Du hast bereits beides: Verbinde das Widget mit den Testdaten und dem Entscheidungsbaum.
Ein Doppelklick zeigt an, wie das Modell die Daten klassifiziert hat (siehe auch Hilfskarte 5).





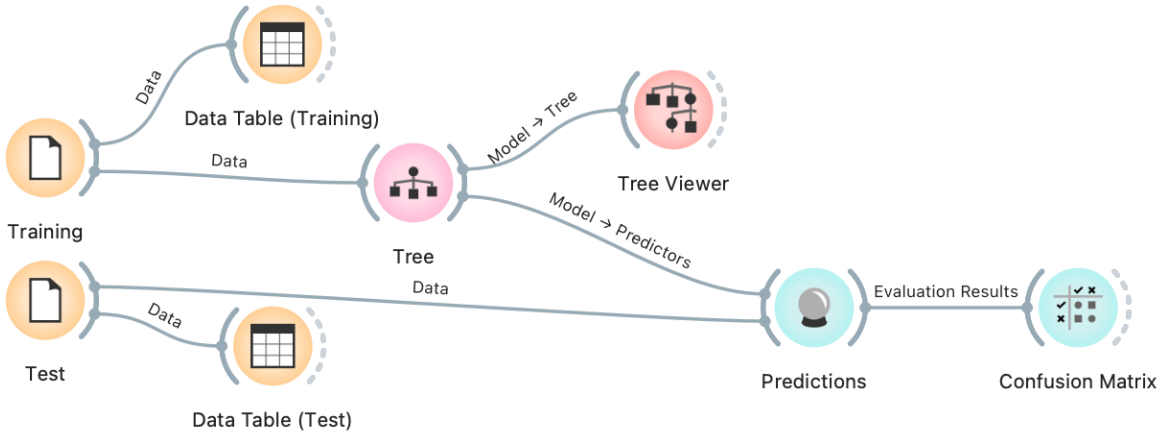
Tatsächliche Klasse

Vorhersage des Entscheidungsbaums mit Wahrscheinlichkeiten für die Klassen beißt (blau) und beißt nicht (rot) an

f) Zum Schluss untersuchst du, wie gut das vom Computer erstellte Modell ist. Bei dieser Einschätzung hilft dir die Konfusionsmatrix.

Ziehe das *Confusion Matrix*-Widget aus dem Bereich *Evaluate* auf die Leinwand (Hilfskarte 6) und verbinde es mit dem *Predictions*-Widget. Ein Doppelklick auf das *Confusion Matrix*-Widget zeigt die Konfusionsmatrix an.

Berechne Genauigkeit, Trefferquote und Präzision deiner Vorhersage gemäß der Formel aus der letzten Stunde. Die Felder der hier angezeigten Konfusionsmatrix entsprechen auch den Feldern aus deinem Arbeitsmaterial von letzter Stunde.



Aufgabe 2:

Jetzt kannst du das Vorgehen des Computers mit deinem eigenen, manuellen Vorgehen vergleichen.

a) Im Spiel aus der letzten Stunde haben wir zunächst Regeln von im Zoo lebenden Äffchen abgeleitet (Trainingsphase) und anschließend das resultierende Modell mit weiteren Äffchen getestet (Testphase). Welche Widgets übernehmen nun diese Aufgaben (Trainings - und Testphase)?

b) Vergleiche den erzeugten Entscheidungsbaum (Doppelklick auf Tree Viewer) mit deiner händisch gefundenen Lösung von letzter Stunde. Beschreibe, ob und worin sich die Entscheidungsbäume unterscheiden.

c) Vergleiche die Konfusionsmatrix und die Metriken deiner händisch gefundenen Lösung mit der des Entscheidungsbaums aus Orange 3 (Doppelklick auf Confusion Matrix). Bewerte den erzeugte Entscheidungsbaum im Vergleich zu deiner händisch erzeugten Lösung.

Aufgabe 3:

Für diese Aufgabe kannst du die Ergebnisse von Aufgabe 1 nutzen. Untersuche die Vorhersagequalität für zwei Baumvarianten:

i) binärer Baum (☑ Induce binary tree) ii) Baum ohne Einschränkungen (☐ Induce binary tree)

a) Vergleiche die beiden erzeugten Entscheidungsbäume. Beschreibe, worin sich die Entscheidungsbäume unterscheiden.

b) Vergleiche die beiden zugehörigen Konfusionsmatrizen (Doppelklick auf Confusion Matrix). Bist du über das Ergebnis überrascht? Wenn ja, warum?

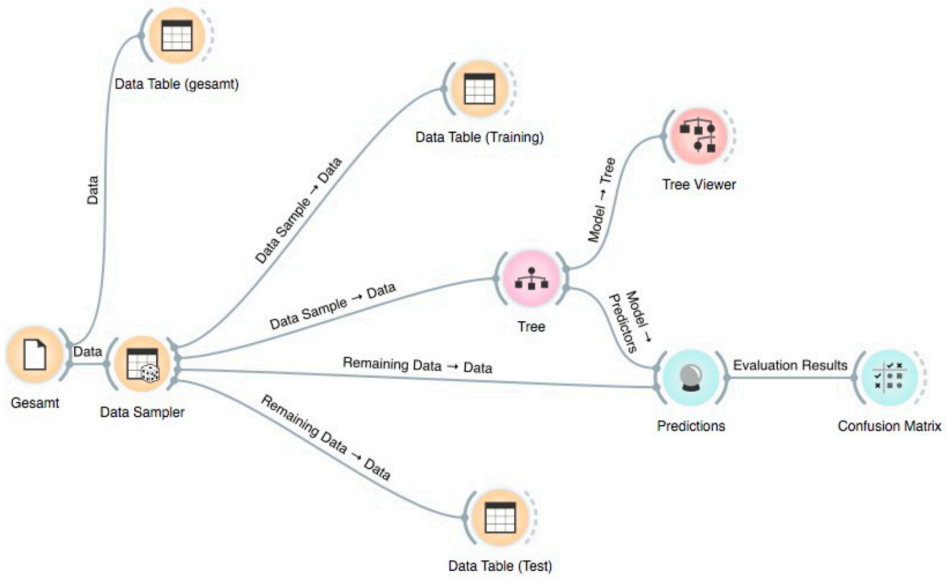
c) Stelle eine Vermutung auf, warum auch Entscheidungsbäume mit weniger Entscheidungsregeln die Vorhersage verbessern könnten!

Aufgabe 4:

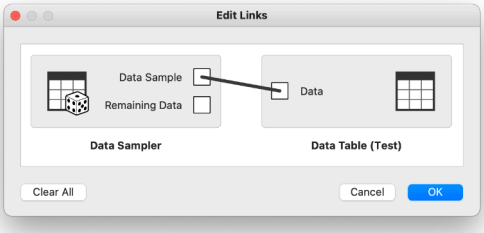
Zum Schluss lernst du, wie sich Test- und Trainingsdaten automatisch aus dem gesamten Datensatz generieren lassen. So wird nicht nur eine zufällige Aufteilung sichergestellt, sondern es spart auch Zeit.

Das Widget *Data Sampler* würfelt die Daten dafür zufällig aus (siehe auch Hilfekarte 7).

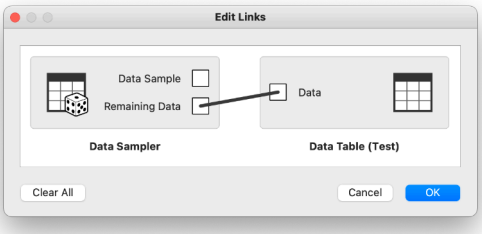
Erstelle ein neues Orange-Projekt für das Äffchenspiel. Als Eingabe steht dabei nur eine Datei (`affen-gesamt.csv`) mit allen Äffchen zur Verfügung. Benenne das *File*-Widget und die *Data Table*-Widgets, wie unten dargestellt. Achte darauf, diesmal nur ein *File*-Widget und das *Data Sampler*-Widget zu verwenden.



Hinweis: Wenn du das Data-Sampler-Widget mit dem nächsten Widget verbindest, steht über dieser Verbindung entweder Data Sample (Trainingsdaten) oder Remaining Data (Testdaten). Um zwischen den beiden zu wechseln, klicke doppelt auf die Verbindung. Die schwarze Linie im sich öffnenden Menü zeigt an, welche Ausgaben des linken Widgets welchen Eingaben des rechten Widgets zugeordnet werden. Die schwarze Linie im Menü kannst du durch einen Doppelklick entfernen und durch Klicken und Ziehen hinzufügen.



Beispiel: *Data Sample* wird als Eingabe für *Data* verwendet.



Beispiel: *Remaining Data* wird als Eingabe für *Data* verwendet.

Variiere den Anteil des Trainingsdatensatzes am gesamten Datensatz, indem du den Regler *Fixed proportion of data* verschiebst. Würfle anschließend den Trainingsdatensatz erneut aus, indem du auf die Schaltfläche Sample Data klickst.

Beobachte dabei die Veränderungen in der Konfusionsmatrix und der Baumdarstellung. Notiere deine Beobachtung!