Einladung zu einem Bewerbungsgespräch: Ja oder nein?

Situation: In einem Betrieb wurden von den vorhandenen Mitarbeiterinnen und Mitarbeitern jeweils folgende Daten erhoben:

Staatsangehörigkeit, Geschlecht, Englischkenntnisse, weitere Sprachen, akademischer Abschluss und Berufserfahrung (über 5 Jahre).

Jedem Mitarbeiter ein Score-Wert (hoch oder niedrig) zugeordnet.
Dem Betrieb liegen nun die Bewerbungsunterlagen von Mara Kur vor.
Da im Unternehmen dringend gute Mitarbeiter gebraucht werden, aber wenig Zeit für zahlreiche Bewerbungsgespräche bleibt, soll vorab schon entschieden werden, ob Mara Kur überhaupt zu einem Vorstellungsgespräch eingeladen werden soll oder nicht.

Wie würden Sie entscheiden?

1. Überlegen Sie zunächst, welche beiden der oben aufgelisteten Merkmale (Staatsangehörigkeit bis Berufserfahrung) für Sie persönlich entscheidend wären, um einen Bewerber zu einem Gespräch einzuladen, und welche Merkmale für Sie unwichtig sind.

Notieren Sie Ihre Wahl und geben Sie eine kurze Begründung dafür an!

Datenanalyse

2. Laden Sie den Datensatz „**bewerbungengrosserdatensatz.xlsx**“ mit dem „File“-Widget in Orange hoch und überprüfen Sie mit dem Widget „Data Table“ Ihren Datensatz auf folgende Kriterien:

Zielvariable (target) festgelegt: ja ☐ , nein ☐ ?

5000 Instanzen vorhanden: ja ☐ , nein ☐ ?

Fehlende Merkmalsausprägungen: ja ☐ , nein ☐ ?

Anzahl der Merkmale: _____ ; Anzahl numerischer Merkmale: _____

Anzahl an Meta-Daten: _____

3. Nehmen Sie im Widget „Distributions“ folgende Einstellungen (rote Pfeile) vor und ergänzen Sie in der Tabelle bei Zeile 1 die Anzahl und den prozentualen Anteil der Mitarbeiter im Unternehmen mit einem niedrigen Score-Wert.

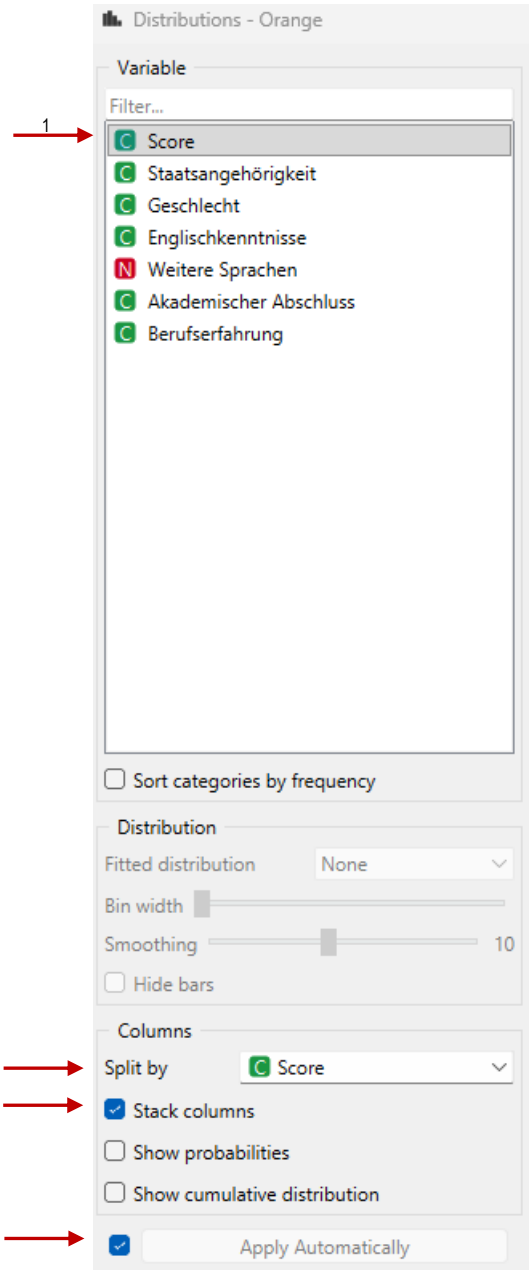


Tabelle:

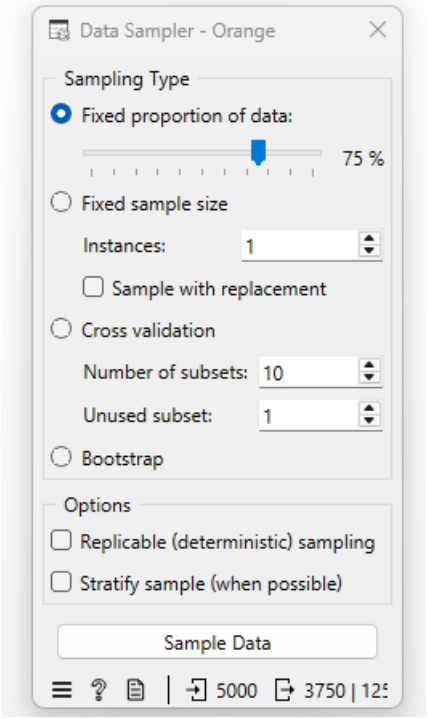
Merkmal	Anzahl Score hoch (**in group**)	Prozentanteil Score hoch (**in group**) **overall**	Anzahl Score niedrig (**in group**)	Prozentanteil Score niedrig (in group) **overall**
Score	1754	35,08 %		
Staats-angehörigkeit (nur dt.)	1736	(35,09 %) 34,72 %	3211	(64,91 %) 64,22 %
Staats-angehötigkeit (österr.)	18	(33,96 %) 0,36 %	35	(66,04 %) 0,70 %
Akadem. Abschluss (ja)	1352	89,77 % 27,04 %	154	
Akadem. Abschluss (nein)	402		3092	88,49 % 61,84 %
Englischkennt-nisse (ja)	496	74,36 % 9,92 %	171	
Englischkennt-nisse (nein)	1258		3075	
Berufserfahrung (über 5 Jahre) (ja)	1304		105	7,45 % 2,10 %
Berufserfahrung (über 5 Jahre) (nein)	450		3141	87,74 % 62,82 %

Ermitteln Sie die fehlenden Tabellenwerte (Häufigkeiten innerhalb der Gruppe (=Merkmalsausprägung) und im Vergleich zur Gesamtzahl an Mitarbeitern durch eine Änderung der Einstellungen im Widget „Distributions“ (bei Pfeil 1).

Nutzen Sie anschließend Ihre Tabellenwerte und bestimmen Sie die beiden wichtigsten Merkmale (inkl. Ausprägung) für die Bewertung eines Mitarbeiters als „hoch“ bzw. „niedrig“. Begründen Sie Ihre Entscheidung!

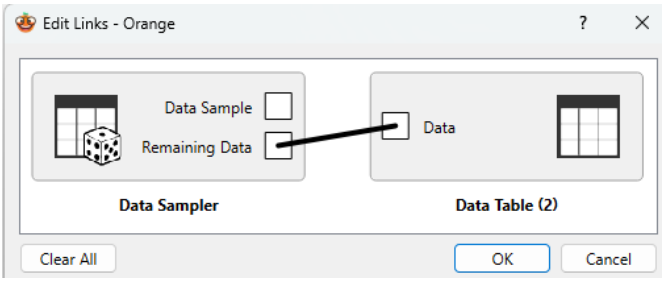
Aufgabenteil 2: Klassifikator erstellen und auswerten

4. Erstellen Sie unter Verwendung des Widgets „Tree“ einen Entscheidungsbaum. Verwenden Sie zum Trainieren Ihres Klassifikators Trainingsdaten, die Sie mit dem Widget „Data Sampler“ aus dem Gesamtdatensatz erzeugen. Bei den Einstellungen im Widget „Data Sampler“ wählen Sie bei „Fixed proportion of data: 75 %“ (siehe Abbildung). Damit werden die 5000 Datenpunkte der Mitarbeiter aufgeteilt in Trainingsdaten (75 % von 5000 = 3750) und Testdaten (25 % von 5000 = 1250).



Beschreiben Sie kurz die Ziele, die mit dem Aufteilen des Datensatzes in Trainings- und Testdaten jeweils erreicht werden sollen.

5. Verbinden Sie 2 weitere „Data Table“-Widgets mit dem Widget „Data Sampler“, um sich die für das Training ausgewählten Daten (= Data Sample) und die für das Testen ausgewählten Daten (= Remaining Data) in Tabellenform anzuschauen. Um die jeweils richtigen Daten an die beiden „Data Table“-Widgets zu schicken, müssen Sie durch Doppelklicken an der Verbindungslinie zweier Widgets, die entsprechende Datenauswahl (siehe Grafik) vornehmen.



(hier: Remaining Data (=Testdaten) ausgewählt)

Trainieren Sie anschließend den Entscheidungsbaum-Klassifikator („Tree“-Widget) mit den ausgewählten Trainingsdaten (=Data Sample).
Tipps: Achten Sie darauf, dass Sie über die Verbindung der Widgets „Data Sampler und „Tree“ die richtigen Daten, d.h. die Trainingsdaten (=Data Sample) schicken.

Beurteilen Sie anschließend die Qualität des Entscheidungsbaums mithilfe der wichtigsten Kennwerte, die Sie über die Widgets „Predicitions“ und „Confusion Matrix“ gewinnen können.

Tipps: Das Widget „Predictions“ benötigt als Eingänge den Testdatensatz (=Remaning Data) und die Ausgabe des Entscheidungsbaums (Models → Predictions).
Die „Confusion Matrix“ benötigt als Dateneingang den Datenausgang von „Predictions“ (Evaluation Results).

5. Für Profis: Beschreiben Sie die Bedeutung der beiden im Widget „Tree“ einstellbaren Parameter: *„Min. number of instances in leaves„* und *„Limit the maximum tree depth to“* im Hinblick auf die Größe und Überschaubarkeit des Entscheidungsbaums sowie auf das Klassifikationsergebnis.

6. Was passiert nun mit Mara Kur? Erhält Sie eine Einladung zum Vorstellungsgespräch?

6.1 Schätzen Sie den Score von Mara Kur anhand ihrer nebenstehenden Merkmalsausprägungen ein!

_____

6.2 Nutzen Sie das Widget „Tree Viewer“, um den Score von Mara Kur abzulesen. Notieren Sie den Score!

_____

6.3 Für Schnelle:
Erstellen Sie anschließend eine Excel-Tabelle mit den Daten von Mara Kur und speichern Sie die Tabelle unter einem geeigneten Namen ab.
Legen Sie unterhalb Ihres bisherigen Workflows einen zweiten an. Verwenden Sie wie oben den gleichen Datensatz (5000 Mitarbeiter) und die gleichen Widgets („File“; „Data Sampler“ (gleiche Aufteilung der Trainings- und Testdaten), „Tree“, „Tree Viewer“, Predictions“) und testen Sie anschließend den entstandenen Entscheidungsbaum mit den Daten von Mara Kur. Notieren Sie den vorausgesagten Score von Mara Kur inklusive der Wahrscheinlichkeit!

_____

Tipp: Um Formatierungsfehler zu vermeiden, kopieren Sie die ersten beiden Zeilen des Mitarbeiterdatensatzes und tragen Sie in der zweiten Zeile die Daten von Mara Kur ein. Beim der Zielvariablen tragen Sie in die Tabelle „?“ ein. Beim Hochladen mit „File“ hilft es manchmal die Zielvariable (hier: Score) nicht als „target“, sondern als „meta“-Attribut zu kennzeichnen.

6.4 Für ganz Schnelle: Vergleichen Sie die beiden Entscheidungsbäume (Widget „Tree Viewer nutzen) und erklären Sie die Unterschiede!