Von einem Unternehmen wurden aus zahlreichen Bewerbungsschreiben von Praktikanten, die sich um einen Praktikumsplatz bewarben, ein Datensatz erstellt. An Informationen wurden beispielsweise Noten, Sprachkenntnisse und Angaben zur Bewerbung erfasst.
Ziel ist es nun, mithilfe von Data Mining herauszufinden, welche Faktoren für den Erhalt eines Praktikumsplatzes eine Rolle spielen und ob sich vorhersagen lässt, ob eine Person ein Praktikumsangebot erhält.

Arbeiten Sie mit der Software Orange Data Mining.

Aufgabenteil 1: Datensatz untersuchen

1. Laden Sie den Datensatz „trainingsdatenpraktikum.csv“ mit dem Widget „File“ in Orange und verschaffen Sie sich über das Widget „Data Table“ einen Überblick über die vorhandenen Daten.
Ermitteln Sie die Gesamtzahl der Datenpunkte (Instanzen), die Anzahl unterschiedlicher Merkmale (features) und die Zahl fehlender Merkmalsausprägungen.

2. Legen Sie unter Angaben von Gründen die Zielvariable fest und stellen Sie diese in Orange ein.

3. Ermitteln Sie mithilfe des Widgets „Distributions“:
- 3 auffällige Merkmale (auffällig: z.B. starke Ungleichverteilung od. unerwartete Verteilung) und notieren Sie jeweils die Häufigkeiten der verschiedenen Merkmalsausprägungen.
- die Zahl der Bewerber mit „0“ und mit „5“ Fehlern in der Bewerbung.
- die Zahl der Bewerber, die eine vollständige Bewerbung abgegeben haben.

4. Formulieren Sie eine begründete Hypothese, welche zwei Merkmale für Firmeninhaber dafür entscheidend sein können, dass Sie Bewerbern einen Praktikumsplatz anbieten.

5. Für Profis: Das Visualisierungswidget „Scatter Plot“ liefert hier – trotz einer Beschränkung auf je 2 Merkmale - nur schwer zu interpretierende Ergebnisse. Erklären Sie diese Tatsache!

Aufgabenteil 2: Klassifikatoren erstellen und vergleichen

6. Untersuchen Sie Ihren Datensatz nach ethisch diskriminierenden Merkmalen und kennzeichnen Sie 2 dieser Merkmale im „file-widget“ als „Meta“-Daten. Dies bewirkt den „Wegfall“ dieser Merkmale aus dem Klassifikator.
Begründen Sie kurz Ihre Wahl!

7. Erstellen Sie zwei Klassifikatoren zur Vorhersage der von Ihnen gewählten Zielvariable.
Tipps: *Nutzen Sie „Test and score“ und „Predictions“.*
Begrenzen Sie im „tree widget“ die Baumtiefe (bei: limit max. tree depth) auf 5, um den Entscheidungsbaum überschaubar zu halten.
Legen Sie beim kNN die Anzahl nächsten Nachbarn auf 4 fest.

7.1 Vergleichen Sie die Korrektklassifikationsraten (KKR od. CA) der beiden Klassifikatoren mit dem Widget „Test & Score“ sowie die Anzahl der korrekt klassifizierten Datenpunkte mit Hilfe des Widgets „Confusion Matrix“. Markieren Sie den „besseren“ Klassifikator!
Tipp: *Das Widget „Confusion Matrix“ nutzt als Eingang die Ausgangsdaten von „Test & Score“.*

7.2 Entnehmen Sie dem Baumdiagramm die beiden wichtigsten Entscheidungskriterien und vergleichen Sie diese mit Ihrer Hypothese aus der Aufgabe 4.

7.3 Entnehmen Sie dem vorliegenden Entscheidungsbaum alle Entscheidungskriterien, die für einen relativ sicheren Erhalt eines Praktikumsplatzes notwendig sind.

7.4 Verändern Sie beim kNN-Modell die Werte von „k“ und notieren Sie die sich daraus ergebenden Unterschiede bei der Korrekt-Klassifikationsrate (= CA-Wert). Markieren Sie den Wert von k, der zum besten Klassifikationsergebnis führt.

7.5 Für Schnelle: Erstellen Sie mit Ihren persönlichen Merkmalen einen Datensatz und testen Sie Ihre Chancen auf den Erhalt eines Praktikumsangebots mithilfe des Entscheidungsbaums.