Use Case – k-Means mit RapidMiner

Der k-Means-Algorithmus gehört zu den Prototyp-basierten Clustering Methoden. Wie der Name schon sagt wird der Mittelwert aller Datenpunkte eines Clusters als Prototyp genommen. Außerdem muss die Anzahl an Clustern mit dem Parameter k zuvor angegeben werden. Diese Punkte bringen sowohl Vor- und Nachteile mit sich, die Sie schon vom Theorieteil kennen sollten. Trotzdem ist der k-Means Algorithmus durch seine einfache Umsetzung ein beliebtes Mittel in der Praxis und relativ weit verbreitet.

Mit welchen Datentypen denken Sie könnte der k-Means Algorithmus Probleme haben?

Dataset importieren

Importieren Sie das „Iris“-Dataset aus dem Ordner Datasets im Materialordner. Die Default-Einstellungen der Import-Funktion können übernommen werden.

1. Ziehen Sie das Dataset in den Prozessbereich.
2. Schauen Sie sich einen Scatterplot von dem Dataset mit den Achsen x: petal_length und y: petal_width an.

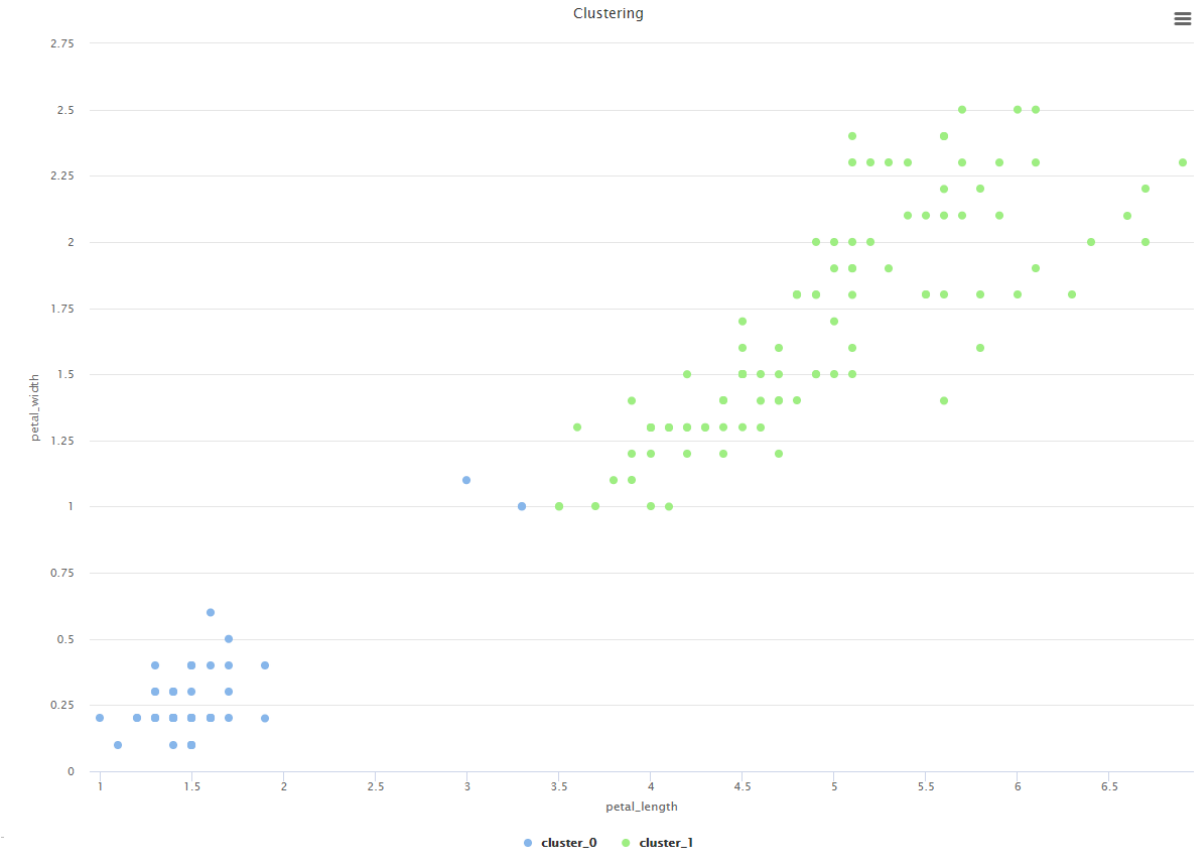
Können Sie abschätzen wie viele Cluster unterschieden werden könnten?

k-Means anwenden

Da bei Unsupervised Learning eigentlich keine Labels vorhanden sind, müssen diese vor der Anwendung entfernt werden.

1. Entfernen Sie das Label mit dem „Select Attributes“-Operator und übernehmen Sie alle numerischen Variablen.
2. Suchen Sie den „k-Means“-Operator und ziehen diesen in den Prozessbereich.
3. Verbinden Sie den „exa“-Knoten des „Select Attributes“-Operators mit dem „exa“-Knoten des „k-Means“-Operators.
4. Wählen Sie als Parameter → „k“: 2 aus.
5. Geben Sie beide „clu“-Knoten aus.
6. Führen Sie den Prozess aus.

Auswertung



Wenn Sie sich nun den gleichen Scatterplot von vorhin anschauen (mit Color = cluster), dann sehen Sie, dass genau zwei Cluster gefunden wurden. Ähnlich wie erwartet wurden beide Cluster voneinander getrennt. Dadurch dass der Feature Raum aus 4 Dimensionen besteht, kann nicht genau gesagt werden, ob die zwei Punkte des grünen Clusters Ausreißer oder wirklich Punkte des blauen Clusters sind.

Erhöhen Sie nun k auf 3 und 4. Schauen Sie sich das Ergebnis erneut an.