Modeling 1: Clustering

Setzen Sie die Parameter des Algorithmus mit Hilfe überwachter oder unüberwachter Evaluationsmethoden. Begründen Sie die Auswahl des Evaluationsalgorithmus.

1. die Parameter des Algorithmus mit Hilfe von überwachten oder unüberwachten Bewertungsmethoden festlegen. Begründen Sie die Wahl des Auswertungsalgorithmus.

Die Parameter des K-Means-Algorithmus, insbesondere die Anzahl der Cluster (n_clusters=4), wurden unüberwacht bestimmt. Ein gängiges Bewertungsverfahren zur Bestimmung der optimalen Clusteranzahl ist die Elbow-Methode oder der Silhouetten-Koeffizient.

- Begründung für die unüberwachte Methode: Da es sich bei K-Means um einen unüberwachten Lernalgorithmus handelt, stehen für die Auswertung keine gelabelten Daten zur Verfügung. Stattdessen wird die interne Kohärenz der Cluster (z.B. minimale Varianz innerhalb eines Clusters) oder die Trennung zwischen den Clustern bewertet.

Die Verwendung einer überwachten Bewertung wäre hier nicht sinnvoll, da es keine bekannten Zielwerte gibt. Theoretisch könnten die Gewichtsklassen als Zielwerte übermittelt werden. Dies ist hier aber nicht gewollt, da die Cluster nicht nur nach Klassen, sondern nach den drei Werten Klasse, Gewicht und Aktivität gebildet werden sollen.

Betrachten Sie die gebildeten Cluster. Wie gut sind sie intuitiv? Welche Informationen über Ihren Datensatz ziehen Sie daraus? Leiten sich weitere Schritte der Datenbereinigung oder der Datenaufbereitung ab?

Die dreidimensionale Grafik stellt die einzelnen Cluster räumlich sehr gut dar. Die Zuordnung der einzelnen Werte ist jedoch schwierig. So kann man kaum erkennen, wie gering das Gewicht der niedrigsten Werte ist, gleichzeitig überdecken sich die Einträge gegenseitig, so dass es schwer zu erkennen ist, ob es Einträge für geringe Aktivität und gleichzeitig hohen Lebensmittelkonsum gibt.

Um diese Hürden zu überwinden, wurden zwei weitere Grafiken mit jeweils nur zwei Dimensionen erstellt. In diesen können die einzelnen Einträge leicht abgelesen werden. Die Zuordnung zu den Clustern kann jedoch teilweise verwirrend erscheinen, da der dritte Wert nicht dargestellt ist. Um die einzelnen Werte gut ablesen zu können, empfiehlt es sich, zuerst die dreidimensionale Grafik zu betrachten und bei Bedarf die beiden zweidimensionalen.

- Intuitive Bewertung der Cluster:
    - Die Cluster scheinen gut voneinander getrennt zu sein. In allen Clustern gibt es eine Art Zentrum mit den meisten Werten. Es gibt jedoch einige Ausreißer. So gibt es Werte von Cluster C in der Mitte von Cluster A. Es handelt sich jedoch um wenige Ausreißer.

Weitere Schritte der Datenaufbereitung:

- Feature-Engineering: Es wurde festgestellt, dass die Aktivität keinen eindeutigen Einfluss auf das Gewicht hat. Ein neues Merkmal in Abhängigkeit von anderen Daten könnte eine bessere Abhängigkeit darstellen.
- Skalierung: Sicherstellen, dass alle Merkmale korrekt standardisiert sind, um Verzerrungen durch unterschiedliche Skalen zu vermeiden.

Können Sie in Ihrem Projektkontext Clustering noch für weitere Zwecke (z.B. Outlier Detection oder Profilerstellung) verwenden? (Dies ist nicht immer der Fall.) Skizzieren Sie ggf. kurz ein mögliches Vorgehen.

Ja, Clustering kann für weitere Zwecke eingesetzt werden:

Outlier Detection:

Punkte, die weit von den Clusterzentren entfernt sind, könnten als Ausreißer betrachtet werden.

Vorgehen:

1. Der Abstand jedes Punktes zu seinem zugehörigen Clusterzentrum wird berechnet.

2. Ein Schwellenwert, beispielsweise basierend auf der Standardabweichung der Abstände, wird festgelegt, um Ausreißer zu identifizieren.
3. Diese Ausreißer werden entfernt oder für die weitere Analyse markiert.

Erstellung von Profilen:

Mit Hilfe von Clustern können spezifische Profile von Datensätzen erstellt werden, z. B. Gruppen von Personen mit ähnlichen Verhaltensweisen oder Gesundheitsmerkmalen.

Vorgehen:

1. Die durchschnittlichen Merkmale jedes Clusters werden analysiert.
2. Die Cluster werden nach den dominierenden Merkmalen benannt (z. B. "hohe Aktivität, niedriger Konsum").
3. Die erstellten Profile werden verwendet, um zielgerichtete Strategien zu entwickeln, z. B. im Bereich der Gesundheitsinterventionen.

Erweiterung

Unter dem Namen „Notebook/Modeling_1_Clustering_Zweiter_Version(Erweitert).ipynb“ findet sich eine erweiterte Clusteranalyse, die auf einer komplexeren Methodik basiert. Diese Analyse berücksichtigt die Dimensionen: *Activity*, *Weight*, *FoodConsumption*, *Sex* und *Age*. Mithilfe dieser detaillierteren Clusterung können ähnliche Daten noch präziser gruppiert und die Einflüsse einzelner Faktoren auf andere besser interpretiert werden. Da eine dreidimensionale Visualisierung nicht alle Dimensionen gleichzeitig darstellen kann, wurden mehrere 3D-Modelle erstellt. Zusätzlich gibt es, wie bei der Standardversion, auch 2D-Modelle für alle möglichen Kombinationen der Dimensionen. Diese Datei stellt eine optionale Erweiterung der ursprünglichen Aufgabe dar und wurde nicht weiter analysiert