1. Definieren Sie für Ihren Datensatz ein oder mehrere Ziele, die Sie mit Hilfe von Dimensionsreduktion der Daten erreichen wollen.

- Besser Übersichtlichkeit durch weniger Variablen
- Sinnvolles Durchführen überprüfen
- Neue Darstellung nur der wichtigsten zwei bis drei Dimensionen ermöglichen

2. Führen Sie mit dem Algorithmus Ihrer Wahl eine Dimensionsreduktion auf Ihren Daten durch.

Wir haben uns für den PCA-Algorithmus entschieden, da wir im Gegensatz zu den Clustering- oder Feature-Selection-Verfahren keine Berührungspunkte mit diesem Algorithmus hatten.

Die Analyse zeigt, dass die ersten 10 Hauptkomponenten bereits 87.43% der Gesamtvarianz erklären. Wir halten diese Varianzabdeckung für ausreichend, um die wichtigsten Muster in den Daten zu erfassen.

3. Setzen Sie ggf. die Parameter des Algorithmus zur Dimensionsreduktion mit Hilfe einer Pipeline.

Wir haben eine Pipeline eingebaut, um einerseits die Zielvariante von mehr als 85% abzudecken und andererseits in Zukunft Erweiterungen mit der Pipeline einbauen zu können.

4. Beschreiben Sie Ihre Ergebnisse. Haben Sie Ihr(e) Ziel(e) erreicht?

Beschreibung:

Das Verfahren der PCA (Principal Component Analysis) reduziert die Dimension der Daten so, dass 85 % der erklärten Varianz erhalten bleiben. Das Ergebnis sind 10 Hauptkomponenten (Principal Components, PC). Die ersten Zeilen der transformierten Daten zeigen die Werte für jede Hauptkomponente (PC1 bis PC10). Diese Hauptkomponenten stellen die reduzierten Dimensionen dar, die die wichtigsten Variationen der Originaldaten erhalten. Der Anteil der erklärten Varianz jeder Hauptkomponente (in Prozent) gibt an, wie viel der Gesamtvarianz durch diese Komponente erklärt wird (PC1: 25,74; PC2: 10,73; PC3: 10,01; usw.). Ein Scatterplot visualisiert die Projektion der Daten auf die ersten beiden Hauptkomponenten (PC1 und PC2). Die verschiedenen Gewichtsklassen (1 bis 4) werden durch unterschiedliche Farben dargestellt, um die Trennung zwischen den Klassen in den Dimensionen zu verdeutlichen.

Ziele:

- Ja, das Ziel, die Dimensionen (Variablen) zu reduzieren und die wichtigsten Informationen (Varianz) zu erhalten, wurde erfolgreich erreicht.

- Es zeigt sich, dass eine Reduktion sinnvoll ist, da einzelne Hauptkomponenten große Teile der Varianz abdecken (bis zu 25,7% in einer Dimension).

- Dabei wurde auch die Reihenfolge der Hauptkomponenten festgelegt, so dass die wichtigsten dargestellt werden können. Die Abbildung zeigt die zwei wichtigsten Komponenten.