Klassifikation

Die Zielvariable für die Klassifikation ist Class, die die Gewichtsklasse darstellt, also ob eine Person Unter/Normal/Übergewichtig oder Adipös ist. Um den besten Algorithmus zur Vorhersage des Zielwertes zu finden, werden SVM, Random Forest, Logistic Regression und XGBoost gegeneinander getestet. Um ein Overfitting zu vermeiden, wird eine Verteilung von Trainingsdaten zu Testdaten von 80%/20% gewählt. Damit stehen sowohl für die Modellbildung und Parameteroptimierung als auch für das Testen der Daten ausreichend Daten zur Verfügung. Um die Genauigkeit der Ergebnisse zu erhöhen, werden diese zusätzlich fünffach kreuzvalidiert.

Für die Auswahl der Merkmale wird die lineare SVM verwendet. Die Merkmale werden automatisch durch SelectFromModel gefiltert. Die ausgewählten Merkmale sind Geschlecht, Alter, Anzahl der täglichen Hauptmahlzeiten, Rauchverhalten, ob Kalorien gezählt werden, Tage mit körperlicher Aktivität und bevorzugtes Verkehrsmittel. Die Modelle mit der besten Performance sind der XGBoost und das Random Forest Modell. Auffällig ist, dass während der XGBoost die Anzahl der täglichen Hauptmahlzeiten als wichtigstes Merkmal erkennt, das Random Forest Modell davon ausgeht, dass das Alter die größte Rolle spielt. Die Gewichtsklasse wird mit dem Random Forest mit einer Wahrscheinlichkeit von 86% richtig vorhergesagt. Die Klassifizierung des Normalgewichts zeigt mit einem F1-Score von 0,92 die beste Leistung. Dies ist sinnvoll, da dies die häufigste Klasse im Datensatz ist (Support: 132). Das Untergewicht hat die höchste Genauigkeit (0.92), aber eine niedrigere Trefferquote (0.73). Dies deutet darauf hin, dass viele Fälle von Untergewicht nicht erkannt werden. Adipositas hat die niedrigste Genauigkeit (0,73), was bedeutet, dass es einige falsch positive Vorhersagen für diese Klasse gibt.

In diesem Fall ist der Recall-Wert wichtiger als der Precision-Wert. Dies Liegt daran, dass falsch-negative Vorhersagen schlimmer sind, da diese Gesundheitsrisiken darstellen. Ein niedriger Recall-Wert würde das Risiko für undiagnostizierte Gesundheitsrisiken erhöhen. Das Modell ist nicht praxistauglich, da die Recall-Werte zu niedrig sind, insbesondere für die kritischen Klassen Adipositas und Untergewicht.