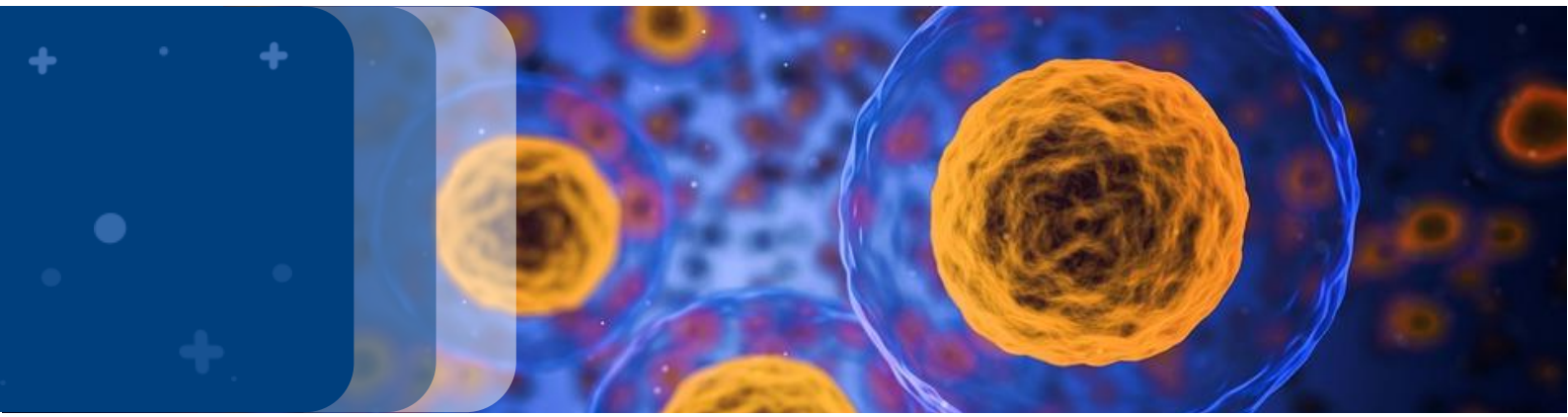
supper&supper



Beurteilung von Brustkrebstumoren mit Hilfe von data iku

Projektziel

Brustkrebs ist die zweithäufigste Krebsart und die erste unter den Frauen weltweit. Im Jahr 2018 wurden mehr als 2 Millionen neue Fälle von Brustkrebs diagnostiziert. Eine Frühdiagnose verbessert die Überlebenschancen drastisch. Die richtige Clusterung von Tumoren in gutartige und bösartige Tumore schützt die Patienten vor der Einnahme unnötiger Medikamente und verringert mögliche Schäden.

Das Projektziel war die Entwicklung eines maschinellen Lernmodells zur Vorhersage der Qualität von Brusttumoren.

Zur Verfügung gestellte Daten

Es wurden zwei verschiedene Datensätze des Krankenhauses Wisconcin mit 570 und 700 Fällen verwendet. Der erste Datensatz basierte ausschließlich auf Daten der Zellenebene. Für jedes Bild wurden zehn Zellmerkmale mit dem entsprechenden Mittelwert, Standardfehler und "worst" (Mittelwert der drei größten Werte) berechnet. Der zweite Datensatz lieferte individuelle Werte von eins bis zehn von Zellattributen und Mitosestadien.

Herausforderungen

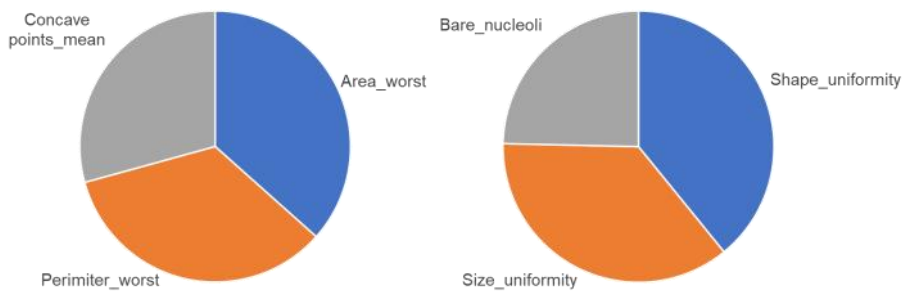
Beide Datensätze enthielten eine heterogene Verteilung von gutartigen und bösartigen Gruppen.

Angewandte Methoden

Aufgrund der Tatsache, dass die Ausgabegruppen bekannt waren, implementierten wir überwachte Lernalgorithmen wie das “random forest” und “logistic regression” Klassifikationsverfahren. Die Genauigkeit wurde als Metrik gewählt, um einen Vergleich zwischen den Leistungen der Algorithmen zu erhalten. Die Methode wurde mit Hilfe der Dataiku-Software durchgeführt.

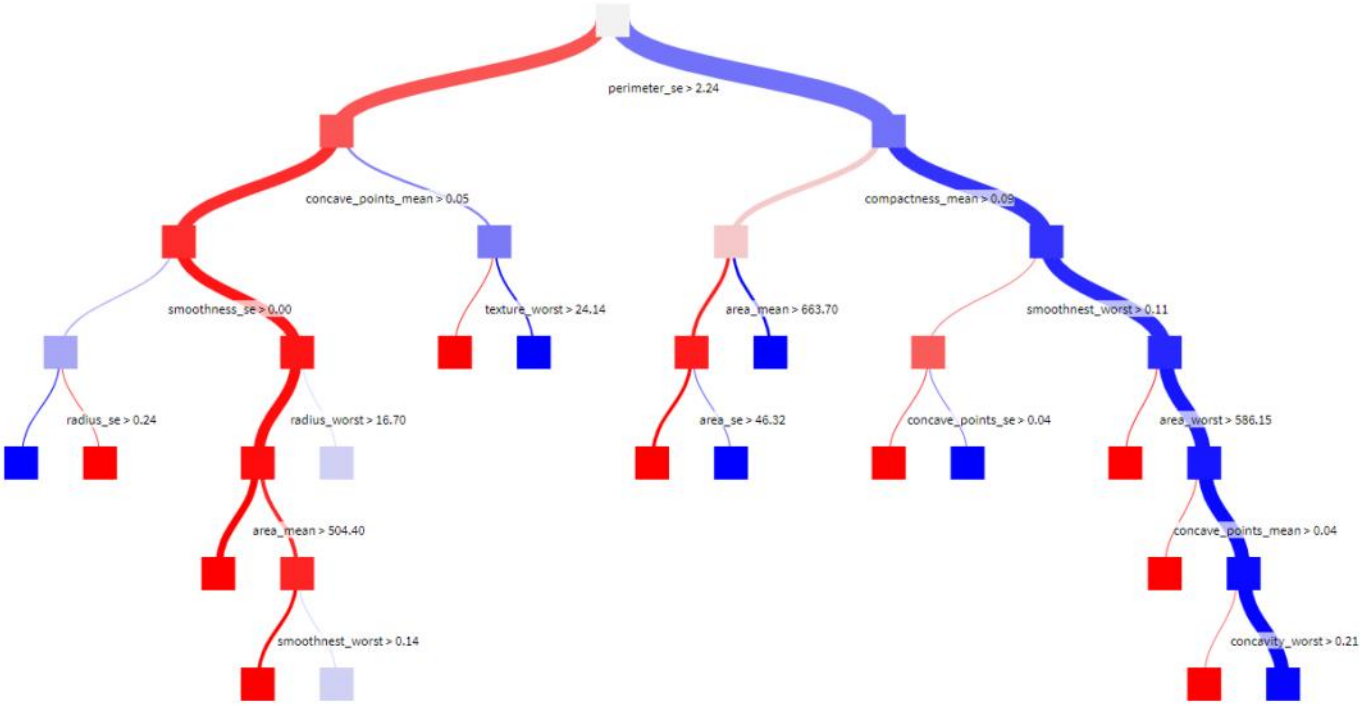
Projektergebnis

Das “random forest” Verfahren erzielte eine Genauigkeit von 99,7 % und 99,3 %.



Variable Relevanz beider Datensätze anhand des “random forest”-Verfahren

Mit Hilfe der “random forest”- Methode haben wir die Bedeutung der verwendeten Variablen beider Datensätze ermittelt. Für den ersten Datensatz waren die Mittelwerte der drei größten Werte entscheidend für die Vorhersage. Es war bemerkenswert, dass das Mitosestadium des zweiten Datensatzes keine Rolle für die Vorhersage spielte.



Entscheidungsbaum des ersten Datensatzes. Rot entspricht den gutartigen Tumoren und Blau den bösartigen. Jeder Hauptpunkt beinhaltet die Wahrscheinlichkeit, dass die Unterpunkte gutartig oder bösartig sind.

Die “logistic regression”- Methode wurde in beiden Fällen mit einer Genauigkeit von 99,6% durchgeführt. Die Konfusion Matrix, basierend auf dem optimierten F1-Score wurden gespeichert. Die falsche Prognose eines gutartigen Tumors anstelle von bösartig wurde höher eingestuft.

	Predicted M	Predicted B	Total
Actually M	37	2	39
Actually B	1	81	82
Total	38	83	121

Der Grenzwert der Konfusion Matrix entsprach der Zahl, ab der die Vorhersage positiv war. Die Werte wurden auf 0,475 und 0,25 gesetzt.

	Predicted M	Predicted B	Total
Actually M	46	0	46
Actually B	4	89	93
Total	50	89	139

Weitere Anwendungen

Die Ergebnisse der Vorhersagen können in der Medizin genutzt werden, um die Einschätzung des Brusttumors mit hoher Genauigkeit zu automatisieren.

Da eine mögliche Therapie vom Risiko des Wiederauftretens von Krebs abhängt, kann der Algorithmus dieses Problem ebenfalls helfen zu lösen.