Mit maschinellem Lernen zum Clinical Decision Support System (CDSS) für Sepsispatienten

Studiengang: MAS Data Science

In der Spital STS AG Thun wird zur Erfassung der Patientendokumentation das Klinikinformationssystem (KIS) CGM Clinical™ eingesetzt. Seit August 2018 ist im KIS das Clinical Decision Support System für Sepsispatienten im Einsatz. Das CDSS besteht aus einem parametrierten Regelwerk, welches primär auf dem SOFA-Score (Sequential Organ Failure Assessment) basiert. Das CDSS soll mit den Techniken aus dem Bereich des maschinellen Lernens verbessert werden.

Zielsetzung

Neben den Optimierungen des bestehenden Clinical Decision Support Systems, soll mit dieser explorativen Arbeit, die Chancen und Risiken der Technologien im Bereich des maschinellen Lernens und dem binären Klassifikationsverfahren aufgezeigt werden.

Methoden

Im Rahmen dieser Arbeit wird zuerst die Systemlandschaft definiert und die Entwicklungsumgebung aufgebaut. Zentrales Element bildet dabei der Apache Spark standalone Kluster.

Für die Datenverarbeitung wird die Spark-MLlib Machine Learning Bibliothek verwendet. Im Bereich der Textanalyse (Natural Language Processing NLP) setzt die Studie auf die Erweiterung Spark-NLP™.

Als Quelle für die Merkmale (Features) dient die Klinikinformationsdatenbank. Dabei werden Sepsis relevante Parameter sowie medizinische Berichte berücksichtigt. Die Grundlage für die Sepsis-Zielvariable, wird durch die interne Abteilung der Kodierung zur Verfügung gestellt. Die Verarbeitung der Daten beginnt mit dem Import aus der SQL-Datenbank in die Spark-Umgebung. Der Prozess der Modellierung wird iterativ nach dem CRSIP-DM Standard (CRoss-Industry Standard Prozess für Data-Mining) durchgeführt. Abschliessend erfolgt die Validierung der Modelle auf einem unabhängigen Datensatz.

Resultate

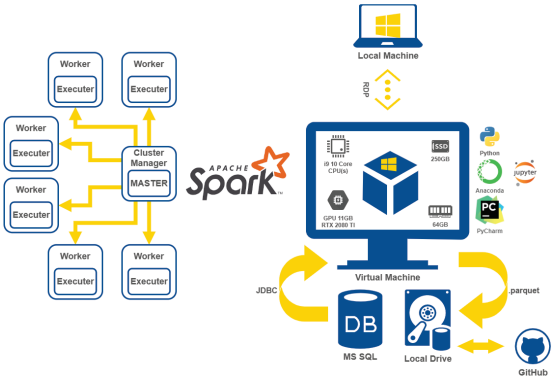
In dieser Studie kristallisierten sich zwei Modelle heraus, welche besonders gute Ergebnisse lieferten. Das erste Modell, basiert auf der Textanalyse der medizinischen Beurteilung, und verwendet für die Klassifikation das Verfahren nach Naives Bayes. Das Zweite Modell beruht auf rein nummerischen Sepsis-Merkmalen. Als Klassifikationsverfahren wird die logistische Regression eingesetzt. Die Güte der Modelle wird mit den folgenden Kenngrössen definiert:

Präzision: 93,8 % (Modell 1) / 82,43 % (Modell 2)
Sensitivität: 85,61 % (Modell 1) / 84,12 % (Modell 2)
Spezifität: 98,61 % (Modell 1) / 82,06 % (Modell 2)
AÛC: 92,11 % (Modell 1) / 83,09 % (Modell 2)

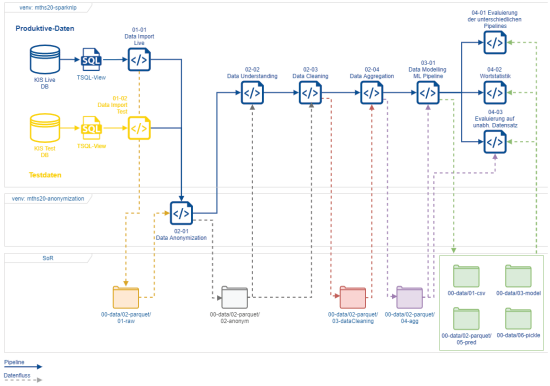
Diskussion

Das bestehende CDSS erreichte in der Post-Interventionsperiode eine Sensitivität von 50,7 % und eine Spezifität von 83,3 % (Marty, 2018). Vergleicht man die Werte mit den Resultaten in dieser Arbeit, dann ist eine deutliche Steigerung der Performance zu erkennen. Es werden mehr Patienten, welche eine Sepsis aufweisen, richtig klassifiziert und auch bei der Gruppe, der nicht Sepsis Patienten, schneidet das Modell besser ab. Vergleicht man die Resultate dieser Studie mit ähnlichen Untersuchungen aus der Literatur, dann kann festgehalten werden, dass sich die Kenngrösse AÛC im ähnlichen Rahmen bewegt.

Nicolas Wiedmer



Übersicht der Sytemlandschaft (Entwicklungsumgebung)



Data Processing Pipeline; Von der Datenbank bis zur Validierung