Detektion von Sepsiserkrankungen in Intensivpatienten

Projektskizze für das Zertifikat "Medical Data Science"

Daniela Vogler

30.04.2022

Motivation und Fragestellung

Eine Sepsis ist eine sehr ernste gesundheitliche Komplikation, insbesondere auf Intensivstationen. Die Sterblichkeitsrate liegt bei schweren Verläufen je nach genauer Abgrenzung der Diagnosen bei um die 50% und höher. Eine frühzeitige Erkennung und damit einhergehend eine rechtzeitige Therapie gelten als wichtigste Mittel zur Verbesserung der Überlebensrate.

In dem geplanten Projekt sollen daher Möglichkeiten untersucht werden, eine Sepsis anhand diverser Blutwerte und Vitalparameter, die üblicherweise im Rahmen von Intensivbehandlungen erhoben werden, mittels prädiktiver Modelle zu erkennen, bevor sie klinisch diagnostiziert wird.

Dem Einsatz von Data Science-Algorithmen zur medizinischen Entscheidungsfindung werden oft Vorbehalte entgegengebracht. Außerdem fehlt es vielerorts an technischen und personellen Ressourcen, um diese Methoden in der täglichen Patientenversorgung einzusetzen. Daher sollen im Rahmen der Projektarbeit auch solche Methoden Berücksichtigung finden, die geeignet sind, nachvollziehbare Entscheidungskriterien zu entwickeln. Solche Kriterien könnten das Vertrauen in die angewandten Verfahren stärken. Zudem könnten sie Eingang in die manuelle Entscheidungsfindung durch Ärzte und Pflegekräfte finden und somit auch ohne die regelmäßige Durchführung von Datenanalysen einen Nutzen in der Patientenversorgung haben.

Aus dieser Zielsetzung ergeben sich folgende zwei Fragestellungen:
1. Welche Entscheidungskriterien hinsichtlich einer Klassifikation von Intensivpatienten in die Kategorien "wird eine Sepsis entwickeln" und "wird keine Sepsis entwickeln" lassen sich mittels einzelner Entscheidungsbäume ableiten?
2. Welches Modell aus dem Bereich der baumbasierten Ensemblemethoden (im Sinne eines methodischen Aufbaus auf Punkt 1) ist am besten geeignet, um aus demografischen Daten, Blutwerten und Vitalparametern eines Intensivpatienten vorherzusagen, ob dieser an einer Sepsis erkranken wird?

Daten

Für das Projekt soll der Kaggle-Datensatz *Dataset.csv*[1] verwendet werden, der unter dem angegebenen Link unter *Open Database License (ODbL)* verfügbar ist. Er basiert auf Daten, die ursprünglich für die PhysioNet Challenge 2019[2] zusammengestellt wurden.

[1] https://www.kaggle.com/binitagiri/hackathon-1feb2022/data?select=Dataset.csv
[2] https://physionet.org/content/challenge-2019/1.0.0/

Der Datensatz umfasst 1.552.210 Beobachtungen zu 40.336 Patienten aus zwei verschiedenen US-amerikanischen Krankenhäusern, von denen 2.932 eine Sepsis entwickelt haben. Er enthält neben einigen Baseline-Angaben zum Fall diverse Vitalparameter und Laborwerte, die in bis zu stündlichen Abständen erfasst wurden. Für eine Auflistung der Variablen siehe Tabelle 1.

Um den Datensatz besser mit der für die Projektarbeit zur Verfügung stehenden Rechenleistung handhaben zu können, soll er auf ein zufälliges Subsample von 10% der Patienten (n = 4.034) eingeschränkt werden.

Methoden

Zur Beantwortung der ersten Fragestellung sollen Entscheidungsbäume angepasst werden.

Zur Bearbeitung der zweiten Fragestellung soll eine geeignete Konfiguration eines Random Forests bestimmt werden. Je nach Vorhersagequalität sind eventuell weitere Ensemblemethoden (zum Beispiel Boostingalgorithmen) heranzuziehen.

Die Daten liegen als longitudinale Daten vor, da die Laborwerte meist täglich und die Vitalparameter teilweise stündlich erfasst wurden. Zur Anwendung der oben beschriebenen baumbasierten Methoden müssen die Messungen daher zunächst geeignet zu einem Datensatz pro Patient zusammengefasst werden. Im Rahmen der Projektarbeit sollen hierfür verschiedene Ansätze evaluiert werden. Mögliche Ansätze sind:
1. Es werden alle Datensätze mit Hour = 1, d. h. die zweite vorliegende Messung, verwendet. Dieses Vorgehen entspricht einer einmaligen Beurteilung der Patienten zu Beginn der Intensivbehandlung bzw. zum erstmöglichen Zeitpunkt.
Hinweis: In dem vorliegenden Datensatz stimmt der Beginn der Messungen nicht mit dem Zeitpunkt der Aufnahme auf der Intensivstation überein, vgl. Variablen 2 und 42 gemäß Tabelle 1. Zudem wird statt der Baseline (Hour = 0) der Zeitpunkt Hour = 1 gewählt, weil es in der ersten Messung extrem viele fehlende Werte gibt. Der Anteil fehlender Werte im Datensatz ist insgesamt hoch, ab Hour = 1 ist jedoch zumindest ein Teil der Variablen (insbesondere die Vitalparameter) schon deutlich besser gefüllt, siehe auch Tabelle 2.
2. Für Sepsispatienten wird die erste mit einem positiven Outcome gelabelte Beobachtung verwendet. Für Nicht-Sepsispatienten wird eine Beobachtung zufällig ausgewählt, da nicht bekannt ist, wann ein solcher Patient der Entwicklung einer Sepsis am nächsten war.
Dieser Ansatz simuliert eine regelmäßige Klassifikation liegender Intensivpatienten anhand ihrer jeweils aktuellen klinischen Parameter.
3. Für jeden Patienten wird mittels linearer Regression die Entwicklung jeder einbezogenen Variable modelliert und der patientenindividuelle Slope an den Algorithmus übergeben. Dabei wird für Nicht-Sepsispatienten der gesamte Beobachtungszeitraum einbezogen und für Sepsispatienten der Zeitraum bis zum ersten positiven Outcome.
Dies entspricht einer regelmäßigen Klassifikation der Patienten anhand aller ihrer bis dahin vorliegenden Parameter.

Alle Ansätze zielen darauf ab, eine Sepsis sechs Stunden vor der ärztlichen Diagnose zu detektieren und somit eine frühzeitige Antibiose zu ermöglichen, denn die Daten wurden bereits sechs Stunden vor dem dokumentierten Ausbruch der Krankheit positiv gelabelt, vgl. Tabelle 1, Variable Nr. 43.

Tabelle 1: Datensatzbeschreibung (Variablen)

Nr.	Variablenname	Datentyp	Beschreibung
1	X	int	Datensatz-ID
2	Hour	int	Stunden seit Beginn der Messungen
3	HR	num	Herzschlag (Schläge/Minute)
4	O2Sat	num	Sauerstoffsättigung (Pulsoxymetrie) (%)
5	Temp	num	Körpertemperatur (°C)
6	SBP	num	Systolischer Blutdruck (mmHg)
7	MAP	num	Mittlerer arterieller Druck (mmHg)
8	DBP	num	Diastolischer Blutdruck (mmHg)
9	Resp	num	Respirationsrate (Atemzüge/Minute)
10	EtCO2	num	endtidaler Kohlendioxidwert (mmHg)
11	BaseExcess	num	Base Excess (mmol/l)
12	HCO3	num	Bicarbonat (mmol/l)
13	FiO2	num	inspiratorische Sauerstofffraktion (%)
14	pH	num	pH-Wert im Blut
15	PaCO2	num	Kohlendioxidpartialdruck (mmHg)
16	SaO2	num	Arterielle Sauerstoffsättigung (%)
17	AST	num	Aspartat-Aminotransferase (IU/l)
18	BUN	num	Blut-Harnstoff-Stickstoff (mg/dl)
19	Alkalinephos	num	Alkalische Phosphatase (IU/l)
20	Calcium	num	Calcium (mg/dl)
21	Chloride	num	Chlorid (mmol/l)
22	Creatinine	num	Kreatinin (mg/dl)
23	Bilirubin_direct	num	Direkts Bilirubin (mg/dl)
24	Glucose	num	Blutzucker (Glukose) (mg/dl)
25	Lactate	num	Laktat (mg/dl)
26	Magnesium	num	Magnesium (mmol/dl)
27	Phosphate	num	Phosphat (mg/dl)
28	Potassium	num	Kalium (mmol/l)
29	Bilirubin_total	num	Gesamtes Bilirubin (mg/dl)
30	TroponinI	num	Troponin I (ng/ml)
31	Hct	num	Hämatokrit (%)
32	Hgb	num	Hämoglobin (g/dl)
33	PTT	num	Partielle Thromboplastinzeit (Sekunden)
34	WBC	num	Leukozytenanzahl (Anzahl*10^3/µl)
35	Fibrinogen	num	Fibrinogenkonzentration (mg/dl)
36	Platelets	num	Thrombozytenanzahl (Anzahl*10^3/µl)
37	Age	num	Alter, Patienten mit 90 Jahren oder älter werden mit 100 kodiert (Jahre)
38	Gender	int	Geschlecht (0 = weiblich, 1 = männlich)
39	Unit1	num	Patient liegt auf einer internistischen Intensivstation (MICU) (0 = falsch, 1 = wahr)
40	Unit2	num	Patient liegt auf einer chirurgischen Intensivstation (SICU) (0 = falsch, 1 = wahr)
41	HospAdmTime	num	Zeit zwischen Krankenhausaufnahme und Verlegung auf die Intensivstation (Stunden)
42	ICULOS	int	Zeit seit Aufnahme auf der Intensivstation (Stunden)
43	SepsisLabel	int	Outcome (1 = Sepsis ist ausgebrochen oder bricht in den nächsten sechs Stunden aus, 0 sonst)
44	Patient_ID	int	Patienten-ID

Tabelle 2: Anzahl fehlender Werte zu Hour = 0 und Hour = 1

X	0	X	0
Hour	0	Hour	0
HR	31581	HR	2313
O2Sat	31817	O2Sat	3163
Temp	33183	Temp	26894
SBP	32307	SBP	4015
MAP	31872	MAP	3153
DBP	34269	DBP	12233
Resp	31908	Resp	4780
EtCO2	40080	EtCO2	38541
BaseExcess	36894	BaseExcess	35054
HCO3	37170	HCO3	36365
FiO2	35829	FiO2	33619
pH	36551	pH	33482
PaCO2	37086	PaCO2	34514
SaO2	39066	SaO2	37042
AST	39935	AST	37773
BUN	36967	BUN	33798
Alkalinephos	39940	Alkalinephos	37788
Calcium	39171	Calcium	34922
Chloride	37068	Chloride	35939
Creatinine	37820	Creatinine	34848
Bilirubin_direct	40285	Bilirubin_direct	40012
Glucose	36657	Glucose	29449
Lactate	39063	Lactate	37360
Magnesium	38944	Magnesium	35407
Phosphate	39463	Phosphate	36968
Potassium	36972	Potassium	32332
Bilirubin_total	40005	Bilirubin_total	37968
TroponinI	40228	TroponinI	38846
Hct	36205	Hct	32983
Hgb	36527	Hgb	33695
PTT	37872	PTT	36670
WBC	37206	WBC	34296
Fibrinogen	39749	Fibrinogen	39616
Platelets	37633	Platelets	34851
Age	0	Age	0
Gender	0	Gender	0
Unit1	15617	Unit1	15617
Unit2	15617	Unit2	15617
HospAdmTime	1	HospAdmTime	1
ICULOS	0	ICULOS	0
SepsisLabel	0	SepsisLabel	0
Patient_ID	0	Patient_ID	0