Healthcare Data Analytics

Einführung

Dr. Michael Strobel, Smart Reporting GmbH

21.03.2022

Organisatorisches

Zeit

Vorlesung: Montag, 10-12 Uhr

Übung: Montag, 12-14

Sprechzeiten: nach der Übung und nach Vereinbarung (m.strobel@posteo.de)

Klausur

Datum: N.N.

Form: wenn möglich mündlich, sonst schriftlich 90 min

Kurzes Intro zu mir

- Aktuell: Vice President Engineering bei der SmartReporting GmbH
- Stationen davor:
  - Principal Software Engineer
  - Freelance Consultant Computer Vision
  - Doktor in Mathematik (Geometrie und Visualisierung) an der TU München
  - Lead Developer einer MINT Visualisierungssoftware: CindyJS
- Geboren und aufgewachsen im Allgäu

Inhalt des Vorlesung

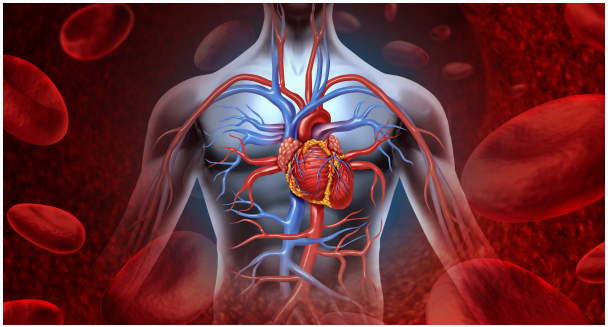
Lernziele

- Methoden der Datenanalyse verstehen und anwenden können
- Einsatzmöglichkeiten von Healthcare Data Analytics im klinischen Umfeld
- Generelle Vorgehensweisen bei Data Analytics und Machine Learning Projekten

Lerninhalte

- Daten verstehen und in Datenstrukturen überführen
- Aufbau von Data Pipelines
- Machine Learning verstehen und anwenden
  - Visualisierung
  - Klassifikation / Regressions / Clustering
  - Decision Trees / Neuronale Netze
  - Training von Modellen
- Anwendungen aus der Praxis
  - Datenstandards
  - Computer Vision
  - Differential Privacy
  - Big Data Tools

Healthcare Data Analytics – Übersicht



freshidea, stock.adobe.com

Datenanalyse und Machine learning in der Praxis

- Erkennung von Mustern in Patientenkohorten
- Interaktive Dashboards zur Informationsgewinnung
- Automatische Diagnose: Erkennung von Tumoren in CT-Scans
- Vorschläge für Therapiemaßnahmen
- Spracherkennung und Sprachsteuerung
- Vorhersage von benötigten Behandungskapazitäten
- ...

Ablauf einen typischen Healthcare Data Analytics Projekts

Definition: Ablauf einen Healthcare Data Analytics (HDA) Projekts

1. Übersicht verschaffen
2. Daten beschaffen und maschinell lesbar machen
3. Daten statistisch auswerten und visualisieren
4. Vorbereitung der Daten für algorithmische Auswertung
5. Selektion der Modelle und Training
6. Beurteilung der Qualität des Modells und Fine Tuning
7. Präsentation der Ergebnisse
8. Deployment, Monitoring und Wartung des Systems

Beispiel: Erkennung von Herzkrankheiten

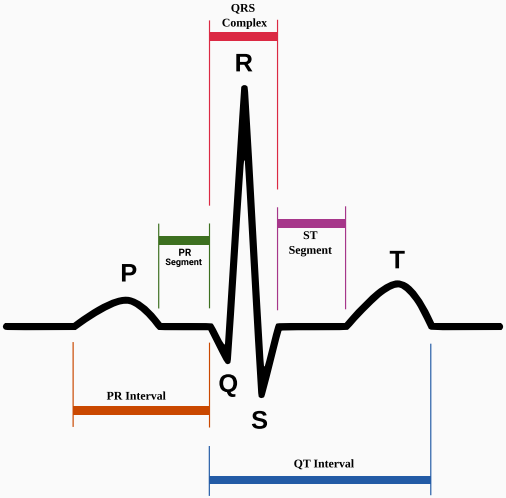
- Als einführendes Beispiel möchte ich mit Ihnen heute ein Machine Learning Projekt durchführen
- Wir erkennen ob ein Patient·in an einer Herzkrankheit leidet oder nicht
- Vorgehen richtet sich nach dem beschriebenen Muster eines HDA Projekts

Übersicht verschaffen: ca. 900 Datensätze

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

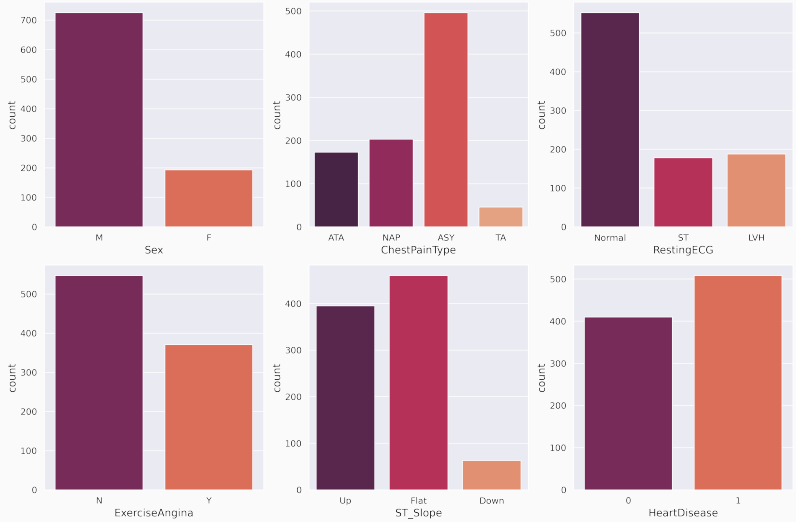
Einige (nicht offensichtliche) Charakteristiken unseres Datensatzes

Feature	Beschreibung	Einheit / Wertebereich
ChestPainType	Art der Brustschmerzen	{TA, ATA, NAP, ASY}
RestingBP	Ruheblutdruck	[mm Hg]
RestingECG	Ruhe-EKG	{Normal, ST, LVH}
ExerciseAngina	Angina bei Belastung	{Y, N}
ST_Slope	Steigung im ST Wert	{Up, Flat, Down}
OldPeak	Abweichung im ST Wert	[-10, 10]
HeartDisease	Output	{1, 0}

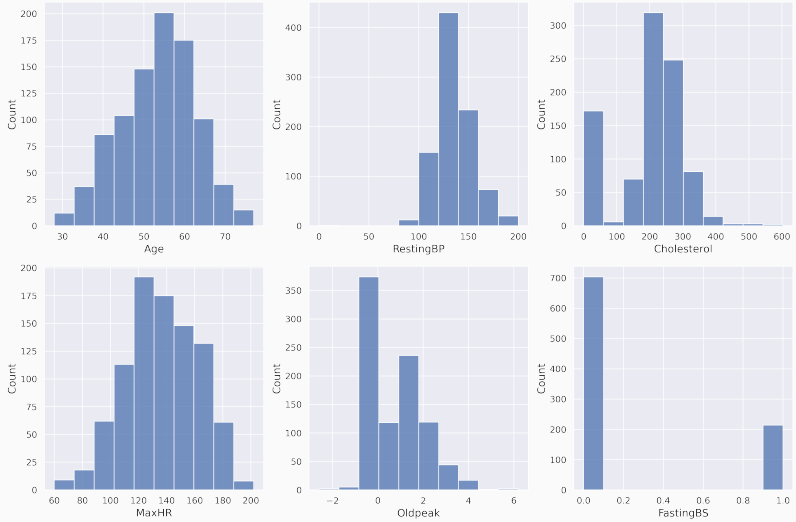


Schematische Darstellung eines EKG

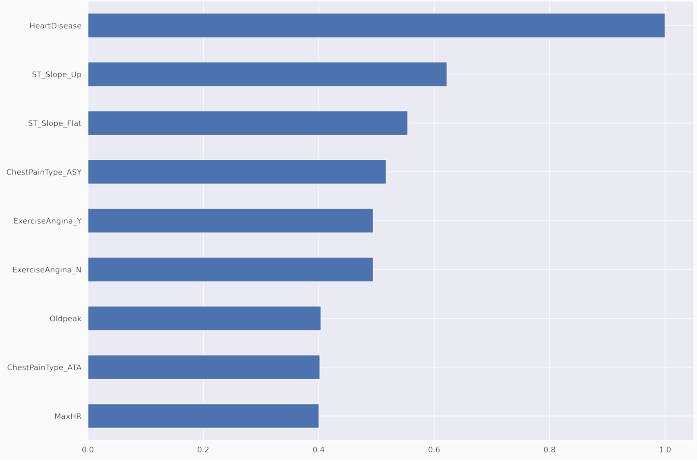
Daten statistisch auswerten und visualisieren



Daten statistisch auswerten und visualisieren



Analyse der Patientendaten und Feature Auswahl – Korrelation Herzkrankheit



Korrelation zwischen Herzkrankheit und einigen Features

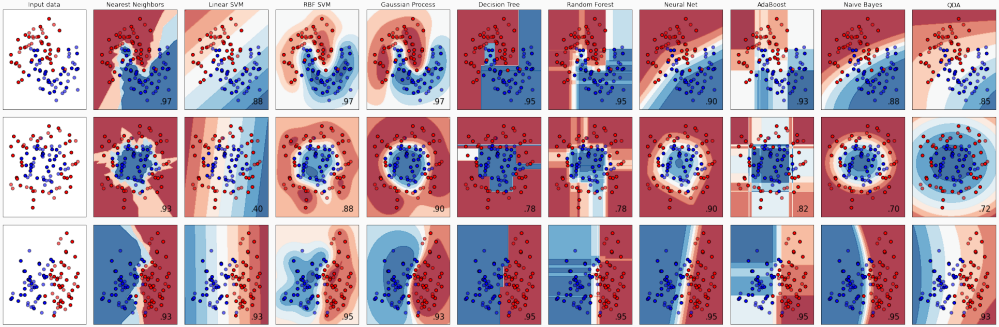
Analyse der Patientendaten und Feature Auswahl – Data Pipelines

Nach der Auswahl von relevanten Features können die **Datapipelines** gebaut werden.

Eine **Datapipeline** ermöglicht es, den Prozess der Datengewinnung, -bereinigung und Transformation zu automatisieren. - Bereinigung der Daten - Umwandlung von Daten (z.B. von Kategorischen Daten in Numerische Daten) - Skalierung von Daten z.B. zwischen 0 und 1

Auswahl des Modells

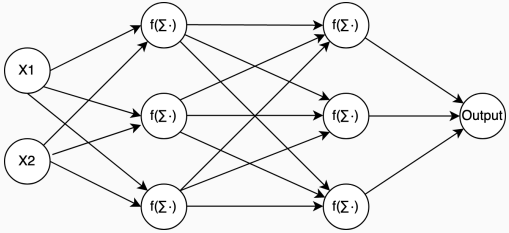
Es gibt eine Vielzahl von Modellen: Decision Trees, Random Forests, Support Vector Machines, ...



Klassifikatoren visualisiert

- Eine Auswahl auf einem kleinen repräsentativen Datensatz stattfinden.
- Wir schauen uns heute die Ergebnisse eines **künstlichen neuronalen Netzes** an.

Künstliches Neuronales Netz



Künstliches Neuronales Netz

- **Trainingsschritt** mit Hilfe von Trainingsdaten und einem **Optimierungsalgorithmus** wird Vorhersageleistung des Netzes verbessert.
- Netzwerkgestaltung und Optimierungsalgorithmen sind aktuelle Forschungsthemen.

Mögliche Ergebnisse einer Binären Klassifikation

- **Richtig positiv (TP)**: Der Patient ist krank, und der Test hat dies richtig angezeigt.
- **Falsch negativ (FN)**: Der Patient ist krank, aber der Test hat ihn fälschlicherweise als gesund eingestuft.
- **Falsch positiv (FP)**: Der Patient ist gesund, aber der Test hat ihn fälschlicherweise als krank eingestuft.
- **Richtig negativ (TN)**: Der Patient ist gesund, und der Test hat dies richtig angezeigt.

Künstliches Neuronales Netz - Ergebnisse

- Wir haben 158 (59 TP + 98 TN) Personen richtig vorhergesagt.
- Bei 27 (18 FP + 9 FN) Vorhersagen lagen wir falsch.

Quellcode: https://github.com/strobelm/heart-failure-prediction

deeperanalytics.be

Unser Fokus: Python, TensorFlow, scikit-learn und pandas -> Übungen

Quellen

- Kardash, M., Elamin, M. S., Mary, D. A. S. G., Whitaker, W., Smith, D. R., Boyle, R., ... & Linden, R. J. (1982). The slope of ST segment/heart rate relationship during exercise in the prediction of severity of coronary artery disease. European heart journal, 3(5), 449-458.
- Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.
- Bishop, C. M. (2006). Pattern recognition. Machine learning.
- fedesoriano. (September 2021). Heart Failure Prediction Dataset.

  https://www.kaggle.com/fedesoriano/heart-failure-prediction.