

## Einführung

Für die Lösung der Klassifikationsaufgabe wurden im Laufe der Lehrveranstaltung 3 Methoden präsentiert und näher betrachtet:

- K-Nearest Neighbors (KNN)
- Multi-layer Perceptron (MLP)
- Support vector machines (SVMs)

und aufgrund der gesammelten Erfahrungen wurden diese 3 zur weiteren Lösung der Aufgabe genutzt. Aus persönlichem Interesse habe ich mit Keras Bibliotheken experimentiert, aber leider auf Grund der mangelnden Erfahrung und Tünnigfähigkeiten, waren damit aufgebaute Klassifikator nicht annähernd so gut wie mit Sklearn aufgestellte Klassifikatoren, deswegen habe ich diese Methode verlassen.

## Features Auswahl

Das gelabelte Dataset beinhaltete 30 Features und über 400 Einträge. Bei so vielen Features bietet sich die Untersuchung der Kreuzkorrelation, um die wenig korrelierte Features zu entfernen, weil sie eher als Rauschen in Daten als Information gesehen werden können. Durch Ausprobieren hat sich der Schwellwert von 0,4 als günstigste geklärt. Wenn Features mit geringeren Korrelationsgrad ausgefiltert wurden, haben alle Verfahren beste Ergebnisse erreicht.

## Diskussion der Modelle

Die untere Tabelle präsentiert die besten erreichten Ergebnisse der Modelauswertung mit bereit zum Training genutzten Dataset (2. Spalte) und mit neuen Daten (3. Spalte).

| Model                        | Traninig ( $R^2$ ) | Test ( $R^2$ ) |
|------------------------------|--------------------|----------------|
| MLPClassifier                | 0.98               | 0.98           |
| Tuned Support Vector Machine | 0.98               | 0.98           |
| Tuned KNN                    | 0.98               | 0.96           |

KNN ist an dieser Stelle nicht günstig zu wählen, da man entweder mit viel größeren Dataset arbeiten musste, um sog. „Curse of Dimensionality“ zu vermeiden oder die Featuresanzahl weiter reduzieren, welche zu Verlusten relevanten Muster in Daten führen könnte.

MLP und SVMs haben vergleichbare Ergebnisse erreicht. Da mir noch Erfahrung mit Tunning fehlt, haben alle eigenen Änderungen die Genauigkeit der beiden Modelle verschlechtert, deswegen wurden die beiden Modelle mit im Praktikum ermittelten Parameter trainiert.

## Das gewählte Model

Obwohl MLP als auch SVMs haben gleichen Ergebnis an Testset erreicht, mag es günstig sein das MLP Modell zur Auswertung der Aufgabe zu nützen, weil sich seine Klassifikationspräzision beim Nachtrainieren mit Testset weiter erhöhen sollte. Der Nachteil von SVMs ist schlechtere Leistungsfähigkeit bei verrauschten, welches bei medizinischen Daten stattfinden kann.