GDS2020\_Bericht\_Flavio\_Müller

Entscheidungsbäume



Flavio Müller

Windisch, 11.09.2020

**Inhaltsverzeichnis**

[1 Executive Summary 3](#_Toc55910174)

[2 Einsatzgebiet 3](#_Toc55910175)

[3 Funktionsweise des Algorithmus 3](#_Toc55910176)

[3.1 Daten Teilen 3](#_Toc55910177)

[3.2 Den Baum wachsen lassen 3](#_Toc55910178)

[3.3 Den Baum zurückschneiden (Pruning) 3](#_Toc55910179)

[3.4 Validating 3](#_Toc55910180)

[3.5 Testing 3](#_Toc55910181)

[4 Entwicklung für die Anwendung 3](#_Toc55910182)

[5 Vor und Nachteile des Algorithmus 4](#_Toc55910183)

[5.1 Vorteile 4](#_Toc55910184)

[5.2 Nachteile 4](#_Toc55910185)

[5.3 Vergleich mit anderen ML-Algorithmen 4](#_Toc55910186)

[5.4 Verbesserungen 4](#_Toc55910187)

[6 Anwendungsgebiete 5](#_Toc55910188)

[6.1 Optimale Anwendungsgebiete 5](#_Toc55910189)

[6.2 Nicht optimale Anwendungsgebiete 5](#_Toc55910190)

[7 Quellenverzeichnis 5](#_Toc55910191)

# Executive Summary

# Einsatzgebiet

* Supervised Learning
* Klassifikation
* Regression

# Funktionsweise des Algorithmus

## Daten Teilen

* Train
* Validate
* Test

## Den Baum wachsen lassen

* Verschiedene Algorithmen
* Target Variable
* Error Algorithmen
* Evtl. einen Algorithmus genau analysieren

## Den Baum zurückschneiden (Pruning)

* Pre Pruning
* Post Pruning

## Validating

## Testing

# Entwicklung für die Anwendung

Schritte aus dem Crisp Data Mining Prozess

A picture containing text

Description automatically generated

# Vor und Nachteile des Algorithmus

## Vorteile

* Leicht nachvollziehbar
* Einfach in der Anwendung
* Prädiktion sind auf klare Regeln zurückzuführen
* Keine Normalisierung der Daten notwendig

## Nachteile

* Durch Schlichtheit, wenig differenziert
* Schlecht bei der Klassifizierung von Komplexen Daten z.B. Bilder
* Anfällig für Bias und Variance

## Vergleich mit anderen ML-Algorithmen

* SVM
* KNN
* DNN

## Verbesserungen

* Bagged Trees
* Boosted Trees
* Random Forest

# Anwendungsgebiete

## Optimale Anwendungsgebiete

* Versicherungswesen -> Einfachheit
* E-Commerce
* Banking
* Meist als Randomforest oder Boosted Trees
* Generell «einfache», für den Menschen verständliche Daten

## Nicht optimale Anwendungsgebiete

* Sehr komplexe Daten (Bilder, Audio usw.)
* Daten welche Menschen schlecht zuordnen können (Pixelwerte von Bildern)

# Quellenverzeichnis