GDS2020\_Bericht\_Flavio\_Müller

Entscheidungsbäume



Figure

Flavio Müller

Windisch, 11.09.2020

**Inhaltsverzeichnis**

[1 Executive Summary 3](#_Toc51183043)

[2 Einsatzgebiet 3](#_Toc51183044)

[3 Funktionsweise des Algorithmus 3](#_Toc51183045)

[3.1 Den Baum wachsen lassen 3](#_Toc51183046)

[3.2 Den Baum zurückschneiden 3](#_Toc51183047)

[4 Entwicklung für die Anwendung 3](#_Toc51183048)

[5 Vor und Nachteile des Algorithmus 3](#_Toc51183049)

[5.1 Vorteile 3](#_Toc51183050)

[5.2 Nachteile 3](#_Toc51183051)

[6 Anwendungsgebiete 3](#_Toc51183052)

[7 Quellenverzeichnis 3](#_Toc51183053)

# Executive Summary

TBD

# Einsatzgebiet

* Klassifikation
* Regression
* *TBD hier noch mehr aufführen vlt. Is Detail gehen ev. Random Forest*

# Funktionsweise des Algorithmus

## Den Baum wachsen lassen

Ein Entscheidungsbaum wächst meistens Anhand von bestimmten mathematischen Regeln, welche zusammen einen Algorithmus bilden. Es ist jedoch auch möglich einen Entscheidungsbaum von Hand wachsen zu lassen. Dies ist jedoch in den meisten Fällen ineffizient. Dennoch werde ich kurz die Grundlegenden Konzepte des Wachsens behandeln.

Zu Beginn brauchen wir wie immer ein Datenset. Dabei sollte darauf geachtet werden, dass das Datenset weder zu klein noch zu gross ist. Abhängig von der Grösse des Datensets, der Problemstellung und der Rechenleistung der Maschine macht es auch Sinn nur ein Teil des Datensets zu benutzen.

* Target Variable
* Split mit möglichst grosser Reinheit
* Reinheit überprüfen
* Schritte wiederhohlen
* Ende wenn:
  + Gewisse Tiefe erreicht (max Layers)
  + Zu wenig daten in einer Node sind (min Records)
  + Weitere Splits die Reinheit nicht mehr signifikant verbessern würden)

## Den Baum zurückschneiden (Pruning)

* Wenn Pre-Pruning (Hyperparameter Tuning) nicht gut genug oder gar nicht verwendet wird kommet es zwingend zu einem Overfit, da der Growing-Algorithmus nicht aufhöhrt die Daten zu spliten
* Das Test-Datenset wird zu gebraucht um das Model zu testen resp. Den Baum zurückzuschneiden Pruning.
* Dabei wird jeder split mit dem Test set analysiert und Splits welche eine schwache relevanz haben (*genau nachschauen was das ist)* werden vom Baum entfernt (der Baum wird zurückgeschnitten).

A screenshot of a cell phone

Description automatically generated

# Entwicklung für die Anwendung

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

## Nachteile

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

# Anwendungsgebiete

* Versicherungswesen -> Einfachheit
* Als Randomforest (Parallel operierende DTs) in der Praxis besser zu gebrauchen evt. Random Forst weiter ausführen

# Quellenverzeichnis