



Inhalt

[Einleitung 3](#_heading=h.gjdgxs)

[Hintergrund 3](#_heading=h.30j0zll)

[Kontext 3](#_heading=h.1fob9te)

[Involvierte Personen 3](#_heading=h.3znysh7)

[Medical-challenge 3](#_heading=h.2et92p0)

[Was ist Uveitis? 3](#_heading=h.tyjcwt)

[Data Cleaning 3](#_heading=h.3dy6vkm)

[Outputvariable 4](#_heading=h.1t3h5sf)

[Methoden zur Berechnung 4](#_heading=h.4d34og8)

[Oversampling/Crosssampling 4](#_heading=h.2s8eyo1)

[K- Fold Crossvalidation 5](#_heading=h.17dp8vu)

[Decision Tree 5](#_heading=h.3rdcrjn)

[Ergebnis 6](#_heading=h.26in1rg)

[Das Package rpart 7](#_heading=h.35nkun2)

[Ergebnis rpart 7](#_heading=h.1ksv4uv)

[Random Forest 8](#_heading=h.44sinio)

[Ergebnis Random Forest 8](#_heading=h.2jxsxqh)

[Vorhersage der Location 8](#_heading=h.z337ya)

[Ergebnis Location Vorhersage 9](#_heading=h.3j2qqm3)

[Location Posterior 9](#_heading=h.1y810tw)

[Idiopathisch 10](#_heading=h.4i7ojhp)

[Ergebnis idiopathisch 10](#_heading=h.2xcytpi)

[Ein Modell ohne idiopathisch 10](#_heading=h.1ci93xb)

[Ein Modell ob idiopathisch oder nicht 10](#_heading=h.3whwml4)

[Löschen von unwichtigen Laborwerten 10](#_heading=h.2bn6wsx)

[Sarkoidose 11](#_heading=h.qsh70q)

[Ergebnis Sarkoidose 11](#_heading=h.3as4poj)

[Zusammenfassung 11](#_heading=h.1pxezwc)

[Bewertung der Methoden 11](#_heading=h.49x2ik5)

[Fazit 12](#_heading=h.2p2csry)

[Anhänge 13](#_heading=h.147n2zr)

# Einleitung

## Hintergrund

### Kontext

Diese Arbeit wurde für die Challenge des Studiengangs DataScience der Fachhochschule Nordwestschweiz erarbeitet. Ziel ist es die Module Supervised und Unsupervised Learning, explorative Datenanalyse, Grundlagen und Vertiefung Datenvisualisierung praktisch zu erlernen. Aus diesem Grund wurde die Challenge: medicalchallenge implementiert.

### Involvierte Personen

|  |  |
| --- | --- |
| Challenge Ausführer | Riccard Nef |
| Challenge Owner | Prof. Dr. Arzu Çöltekin, Dr. Fernando Benites |
| Auftraggeber | Prof.Dr.Nida Sen and Dr. Shilpa Kodat |
| Zeitraum | HS 2020(20.9.2020-24.1.2021) |

### Medical-challenge

Das Ziel der Challenge ist die Erstellung eines mathematischen Models für die Vorhersage der Diagnose des Patienten. Zu diesem Anlass wurde eine [Liste](http://data/uveitis_fulldata_v2.xlsx) von 1075 untersuchten Patienten zusammengestellt, die an Uveitis erkrankten. Diese Liste ist anonymisiert und enthält folgende Informationen:

1. Geschlecht und Rasse
2. Diagnose
3. Location
4. Über 30 verschieden Laborwerte
5. DNA-Daten

Die genaue Beschreibung der Challenge finden Sie [hier](http://data/projectdescribtion_cml1_HS2020_%20FHNW.pdf).

### Was ist Uveitis?

Die Uveitis beschreibt eine Entzündung der mittleren Augenhaut, der Ueva. Diese Entzündung kann in verschiedenen Positionen im Auge entstehen, diese werden Anterior, Intermedia, Posterior, Pan oder Skleritis genannt. Die Ursache dieser Krankheit ist in den meisten Fällen nicht erkennbar, was in aus medizinischer Sicht idiopathisch genannt wird. Allerdings gibt es auch klare Krankheitsfälle. Zum Beispiel werden c.a 30-50% aller Menschen, die an Morbus Bechterew leiden, mindestens eine Uveitis durchleiden. Andere Begleitkrankheiten können Tuberkulose, Masquerade, Syphilis und vieles mehr sein. Zusammengefasst gibt es so 8 Kategorien, wonach sich die Diagnosen einteilen lassen: Idiopathic, systemic, wds, nonneoplastic masquerade, neoplastice masquerade, infectious, not uveits and Scleritis.

Aus dieser grossen Unsicherheit ist es für Medizinerinnen ein grosses Anliegen eine Verbesserung in der Diagnostizierbarkeit von Uveitis zu erreichen.

# Data Cleaning

Die Liste mit allen Informationen befindet sich in einer Excel Tabelle. Für eine fehlerfreie Berechnung ist es unabdingbar die Daten auf Unstimmigkeiten und Fehler zu überprüfen. Diese Überprüfung hat ergeben, dass folgende zwingende Bearbeitungsschritte notwendig sind:

1. Alle geschriebenen Informationen müssen klein geschrieben werden und auf Fehler überprüft werden.
2. Die Namen der Spalten müssen geändert werden.
3. Die DNA- Werte werden nicht miteinbezogen also müssen sie gelöscht werden.
4. Die Laborwerte haben unterschiedliche Masseinheiten, diese müssen angepasst werden.
5. Die Laborwerte wurden von unterschiedlichen Tests erhoben, was eine andere Bewertung des Laborwertes erfordert. Jede Testart gibt die höchst und mindestgrenze des normalen Laborwerttests an. So kann man die Laborwerte mit folgender Formel standardisieren.

So hat jeder Wert, der sich innerhalb der Range befindet, den Wert von 0 bis 1. Alle Werte, die das Maximum überschreiten haben einen Wert über 1 und alle Werte, die das Minimum unterschreiten haben einen negativen Wert.

Das Datacleaning wurde in der Programmierspreche Python durchgeführt und befindet sich in folgendem Jupyter-Notebook [Datei](http://nordstream.ipynb). Diese Datei ist auch als HTML verfügbar.

## Outputvariable

Das Ziel ist mit den Laborwerten die Diagnose vorherzusagen. Doch es gibt 3 Spalten, die diese Diagnose beschreiben. Die EHR-Diagnose konnte in Absprache ignoriert werden. Dann gibt es noch die Special Diagnosis, was die echte Diagnose darstellt. Allerdings auch eine Kategorielle Diagnose, welche die Special Diagnosis gewissermassen zusammenfasst. Es wird versucht die Special Diagnosis vorherzusagen, es ist aber klar, dass die Vorhersage der kategoriellen Diagnose immer besser sein wird als die Vorhersage der Special Diagnosis. Denn je weniger verschiedene Klassen es vorauszusagen gilt, desto besser wird das Model.

# Methoden zur Berechnung

Alle Vorhersagemodelle wurden in der Programmiersprache R verfasst und finden sich in folgendem [RNotebook](http://uveitscalculation.rmd). Diese Datei ist auch als HTML verfügbar.

## Oversampling/Crosssampling

Um die kategorielle Diagnose vorherzusagen ist es wichtig eine möglichst gleiche Verteilung der Output Variable zu gewährleisten. Man sieht im unteren Bild, das die Verteilung sehr unterschiedlich ist, man nennt dies Imbalance data. Das grösste Problem bei Imbalance data ist, dass im schlimmsten Fall das Modell nur noch idiopathisch vorhersagt und in 50% richtig liegt. Vor allem die Berechnung mit Random Forest hat gezeigt, dass dies durchaus möglich ist. Wenn bei einer Random Forest Modellberechnung sehr viele Bäume genutzt werden, passiert genau dies.

Ein Lösungsansatz zu diesem Problem ist das sogenannte oversampling. Hier wird mit verschiedensten Methoden synthetische daten erschaffen, sodass die Verteilung gleichmässig wird. Die zusätzlichen Daten sind nicht einfach Kopien der daten, sondern wurden durch eine k-neares neighbor Algorithmus entworfen. So wurde die Anzahl der Patienten von 1075 auf 2954 erhöht.

Zusätzlich wurde ein Datensatz erstellt, der mit crossampling implementiert wurde. Dieser Datensatz hat 1933 Patienten daten. Hier wurde eine Kombination aus oversampling und undersamling verwendet.

Dieser Prozess wurde mit der SMOTE Methode nach dem Vorbild eines [Webartikels](https://towardsdatascience.com/imbalanced-data-oversampling-using-genetic-crossover-operators-bb581150a9a8) durchgeführt.

## K- Fold Crossvalidation

Die K- Fold Crossvalidation wurde genutzt, um einen groben Überblick über die Daten zu erhalten. In dieser Methode wird der Datensatz in 10 Teile geteilt, welche danach einzeln benutzt werden, um ein Modell davon zu erstellen. Die Ergebnisse der Modelle werden dann miteinander verglichen und sollten möglichst gleich sein. Wenn dies nicht der Fall ist, heisst dass das die Daten zu zerstreut sind für ein mögliches Modell.

#### Ergebnis

Das Ergebnis der K-Fold Crossvalidation ist, dass der Fehlerwert des Models bei c.a 55% sehr hoch war. Die Standartabweichung der Werte lag allerdings bei nur 2% was auf eine gewisse Struktur der Daten hinweist.

|  |  |  |  |
| --- | --- | --- | --- |
|  | normal Data | oversampled Data | crosssampled Data |
| error rate mean | 0.556 | 0.666 | 0.666 |
| error rate max | 0.59 | 0.7 | 0.7 |
| error rate min | 0.51 | 0.63 | 0.63 |
| error rate std | 0.026 | 0.22 | 0.22 |

Es zeigt sich das die oversampleten daten ein signifikant schlechteres Ergebnis erzielten. Das Ergebnis der Crossampled daten ist genau so schlecht wie des oversampled daten. Dass heisst das die gesampleten daten auf den ersten Blick kein besseres Ergebnis erzielen und dass es nicht darauf ankommt, ob die synthetischen daten mit oversampled oder crossampled erstellt wurden

## Decision TreeDecision Tree and Entropy algorithm. - ZhengTianyu's Blog

Ein Decision Tree (DT) ist ein mathematisches Modell, dass versucht eine Outputklasse zu erkennen. Er gehört zur Klasse der supervised learning Algorithm. Die Vorteile des DT sind eine gute Erklärbarkeit. Denn es gibt viele Algorithmen, deren Berechnung nicht nachvollziehbar sind. Der DT hingegen kann sogar visuell dargestellt werden. Wie man in diesem Beispiel sieht:

Ein weiterer Vorteil ist, dass der DT NAN values handeln kann. Der DT ist einer der wenigen Algorithmen, der mit nichtexistierenden Werten umgehen kann. Allerdings war es nur möglich in der Programmiersprache R diese nichtexistierenden Werte miteinzubeziehen. Was auch der Hauptgrund der für die Verwendung dieser Programmiersprache war.

Weitere detailliertere Informationen zum Decision Tree finden Sie [hier](https://www.upgrad.com/blog/gini-index-for-decision-trees/).

### Ergebnis

#### Kategorische Werte

Der Datensatz wird in 2 Teile unterteilt, mit einer Verteilung von 1:5. Diese Teile werden train und test genannt. Es ist sehr wichtig, dass das Ergebnis des Modelles auf anderen Daten getestet wird, als auf es entwickelt worden ist. Nach der Teilung und dem Erstellen des Modelles wird dieses nun auf die Testdaten angewandt und es wird geschaut wie viele richtig sind(accuray) oder wie viele falsch sind (error rate). Dieser Wert zeigt die Qualität des Modelles auf, welche man gegeneinander vergleichen kann.

|  |  |  |  |
| --- | --- | --- | --- |
|  | normal Data | Oversampled Data | Crossfitted Data |
| error rate | 0.4148 | 0.4155 | 0.35 |
| Used labs (labs are ordered by importance) | - location  - abn\_od\_haze  -. rv | - location  - abn\_od\_haze  - rv  - AC Abn Os Cells  - MCHC (Blood)  - Angiotensin Conv#Enzyme  - WBC (Blood) | - abn\_od\_haze  - location  - Vit Abn Os Cells  - Vit Abn Od Cells  - Vit Abn Os Haze  - AC Abn Os Cells  - Lymphocytes %  - HCV (HepC) Ab  - Antinuclear Antibody  - Beta-2-Microglobulin  - Complement C3  - RBC  - AC Abn Od Cells  - C-Reactive Protei  - HBs (HepB surface) Ag |
| Predictions correct | Idiopathic (82/106)  Infectious(7/17)  Neo. Masq(3/4)  Nonneo Masq(1/23)  Not uveitis(0/0)  Sclerits(0/0)  Systemic(14/38)  Wds(3/21) | Idiopathic (67/107)  Infectious (0/17)  Nonneo Masq(1/4)  Systemic (13/38)  Wds(0/25) | Idiopathic (0/107)  Infectious (1/17)  Nonneo Masq(1/4)  Systemic (1/38)  Wds(0/25) |
| Source | [Confusion Matrix 1](http://data/CMDT-normal.png) | [Confusion Matrix 2](http://data/CMDT-oversampled.png)  Legend:  0. idiopathic  1. Infectious  2. Neopl. Masq  3. Nonneo Masq  4. systemic  5. wds | [Confusion Matrix 3](http://data/CMDT-crosssampled.png)  Legend:  0. idiopathic  1. Infectious  2. Neopl. Masq  3. Nonneo Masq  4. systemic  5. wds |

Ein normaler Decision Tree hat eine Fehlerquote von 41%, was weder gut noch schlecht ist. Es hat sich gezeigt, dass das Model nur 3 Laborwerte nützt, um die Kategorie zu bestimmen. Der Erfolg des Modells ist hauptsächlich mit der hohen Anzahl der idiopathisch Fälle zu erklären. Denn mit der Vorhersage der anderen Kategorien ist das Modell sehr schlecht.

Auch hier schneiden die Over und Crosssampled Daten schlechter ab als das Modell mit den normalen Daten. Der Error rate der crossampleten Daten ist zwar gering. Wenn man allerdings die genaue Vorhersage anschaut, sieht man, dass das Modell fast nichts richtig vorhergesagt hat.

Ein spannender Fakt ist hier, dass die verwendeten Laborwerte ähnlich sind. Es fällt auf, dass die Augenwerte der Abnomalen Zellen und der Hazewert einen grossen Einfluss haben kann.

#### Vorhersage der Special Diagniose

|  |  |
| --- | --- |
|  | normal Data |
| error rate | 0.42 |
| Used labs (labs are ordered by importance) | - location  - rv  - Lymphocytes  - abn\_os\_cells |
| Predictions summary | - idiopathic\_anterior(31/32)  - idiopathic\_panuveitis(21/21)  - idiopathic\_posterior(6/11)  - idiopathic\_scleritis(10/10)  - hla\_b27(4/4)  - wds(10/18)(most wrong are noneopl Masq)  - vka(0/9)(most wrong are idiopathic\_panuveits)  - toxoplasmosis (5/6)  - nonneoplastic\_masquerade(8/23)(9 of them are wds)  - pars\_planitis(29/29) |
| Source | [Confusion Matrix 4.1](http://data/CMDT-diagnosis1.PNG)  [Confusion Matrix 4.2](http://data/CMDT-CMDT-diagnosis2.PNG) |

Es zeigt sich, dass die Vorhersage einigermassen funktioniert. Die Fehlerquote liegt nur bei 0.4, was für eine Klassifikation von 27 Klassen eigentlich sehr gut ist. Bei genauer Betrachtung der Qualität der Vorhersage sieht man, allerdings dass auch dieser Erfolg nur der idiopathic zu verdanken ist. Das Modell ist sehr gut im Erkennen von idiopathischen Fällen, für den Rest ist das Modell allerdings sehr schlecht und erkennt fast nichts. Eine bemerkenswerte Ausnahme spielt hier die Pars Planitis. Welche zu 100% richtig erkannt wurde.

## Das Package rpart

Di obigen Ergebnisse wurden mit dem R package tree generiert. Da die Darstellung des Decision Trees packages sehr schlicht ist und nur begrenzt veränderbar ist wurde ein alternatives package versucht, das rpart package. Es wurde für die normalen, Oversampleten und Crosssampelten Daten ein Decision Tree mit diesem package erstellt und die error rata verglichen.

### Ergebnis rpart

Die Ergebnisse des rpart package sind:

|  |  |
| --- | --- |
| normale Daten: | 0.52 |
| oversampled data | 0.78 |
| crossampled data | 0.99 |

Es hat sich gezeigt, dass das package rpart um einiges schlecht in der Vorhersage der kategoriellen Diagnose ist, als das package tree. Darum werden hier nur die Ergebnisse des tree packages detailiert beschrieben. Die error Rate der crossampled data liegt bei 99%. Dies ist auch ein weiter Hinweis, dass die oversampled und crossampled data keine Verbesserung des Modelles bewirken.

## Random Forest

Bei der Erstellung der Decion Trees ist es aufgefallen, dass das Ergebnis variieren kann. Je nach verwendeten Daten verändert sich der Baum. Dies liegt an dem Teilen der Daten, welche auf dem Zufallsprinzip beruht. Dies soll der Algorithmus Random Forest besser machen. Denn dieser verwendet eine Vielzahl von zufällig generierten Bäumen. Leider kann der Wald nicht visualisiert werden, es ist allerdings möglich eine klarere Übersicht über die Wichtigkeit der Labortests zu gewinnen.

Das Random Forest Modell kann nicht min «Nicht existierenden» Werten arbeiten, also mussten diese durch den Mittelwerde ersetz werden. Insgesamt wurden 4032 Werte ersetzt.

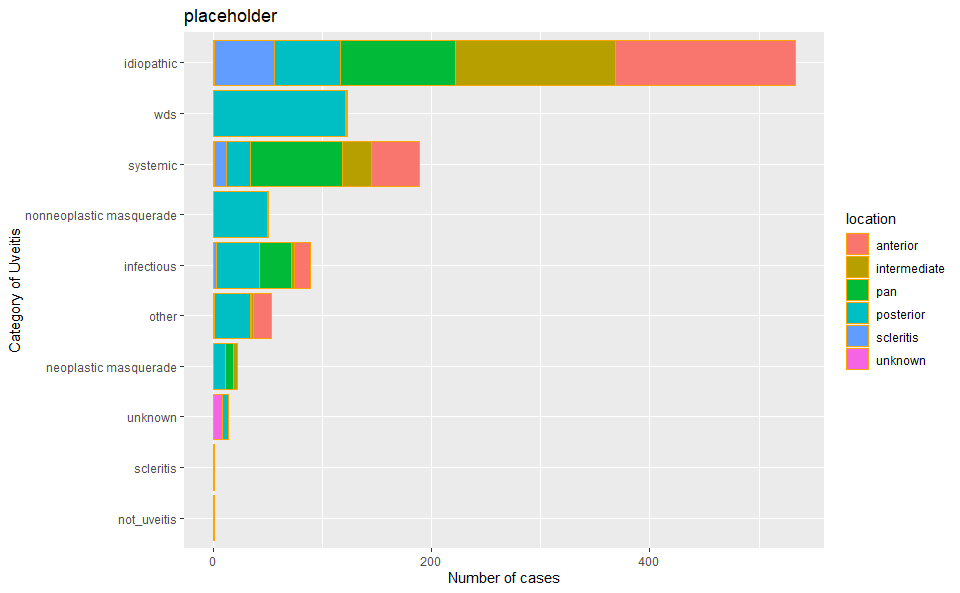
### Ergebnis Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
|  | Kategorielle Diagnose | Special Diagnose | Oversampled data |
| error rate | 0.49 | 0.78 | 0.88 |
| Used labs (labs are ordered by importance) | - location  - ac\_abn\_od\_cells  - ac\_abn\_os\_cells | - location  - abn\_od\_haze | - abn\_od\_haze  - location |
| Source | [Confusion Matrix 5](http://data/CMRF-normal.PNG) | [Confusion Matrix 6](http://data/CMRF-diagnosis.PNG)  [Confusion Matrix 7](http://data/CMRF-diagnosis2.PNG) | [Confusion Matrix 8](http://data/CMRF-oversampled.PNG) |

Die Ergebnisse zeigen, dass der Random Forest Algorithmus wesentlich schlechtere Resultate liefert als der Decision Tree. In der Confusions Matrix des kategoriellen Diagnose sieht man, dass das Modell einfach alle Klassen als Idiopathic klassifiziert und dann mit 50% richtig liegt. Dies ist nicht ein gelungenes Modell und die error rate der Special Diagnose und der oversampled data sind noch schlecht.

Der Algorithmus scheint Probleme mit der unterschiedlichen Verteilung der Output Variable zu haben.

## Vorhersage der Location

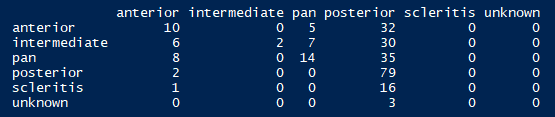
6Der Laborwert der Location scheint in den berechneten Modellen eine zentrale Rolle zu spielen. Er ist immer mit Abstand der wichtiges Laborwert. Der Location Test geschieht anhand der Arztdiagnose und kann mit einer speziellen Kamera Technik meistens klar bestimmt werden. 

### Ergebnis Location Vorhersage

Die grafische Darstellung der Verteilung der Location zeigt auf, dass je nach Kategorie verschiedene Locations-verteilungen auffinden. Z.B ist wds und eine maskierte Uveits fast nur in der Location posterior vertreten oder eine unbekannte Location führt in vielen Fällen auch zu einer unbekannten Kategorie.

Es wird nun untersucht ob nur mit den Laborwerten die Location vorhergesagt werden kann.

Die Fehler Rate liegt bei 0.58. Dies bedeutet, dass es besser ist mit der Location die kategorielle Diagnose vorherzusagen, also ohne. Denn das Modell der Location Vorhersage ist um fast 0.2 schlechter als die der kategoriellen Diagnose.

Das Modell scheint die meisten Klassen posterior zuzuweisen und obwohl der Unterschied der Anzahl der Klassen gering ist.



### Location Posterior

Aus dem Ergebnis der obigen Untersuchung wird nun untersucht, ob die Fehlerrate sich verändert, wenn der Datensatz nur noch aus Patienten mit der Location Posterior besteht. Dieser neue Datensatz wurde benutzt, um eine kategorielle Diagnose Vorhersage mit dem Decision Tree zu machen.

#### Ergebnis Location Posterior

Die Fehlerrate des Modelles liegt bei 0.32. Dies ist um einiges besser als die vorherigen Modelle. Dies weisst darauf hin, dass es sinnvoll sein kann die Vorhersagemodelle für die Location separat zu erstellen. In Posterior ist es einfacher eine Vorhersage zu tätigen als in anterior oder intermediate.

Das Modell hat folgende Laborwerte für die Location Posterior verwendet:

1. rv
2. complement\_c3
3. Hematocrit
4. Calcium
5. Mchc
6. Mcv
7. ld

Wenn der Patient Granulome hat, ist es nach dem Modell sehr wahrscheinlich idiopathisch. Es fällt auch auf, dass die Entscheidungen des Decision Trees, für Laborwerte zwischen 0 und 1 liegen. Dies bedeutet das es nicht nur wichtig ist, dass ein Patient in der Range des Labortestes ist, sondern auch wie hoch der effektive Wert ist.

# Idiopathisch

Idiopathisch ist der medizinische Ausdruck der Unwissenheit. Also in eine Idiopatische Diagnose eine unbekannte Diagnose. Die bisherigen Modelle haben hauptsächlich Idiopathische Diagnosen richtig vorhergesagt. Was eigentlich ein Wiederspruch in sich selbst ist. Denn man bestimmt etwas Unbestimmtes und es ist auch fragwürdig was diese Vorhersage für die Behandlung für Konsequenzen hat. So wurden folgende Berechnungen durchgeführt:

- Vorhersage ob idiopathisch oder nicht

- Fillern des Datensatzes nach nicht idiopathischen Diagnosen

## Ergebnis idiopathisch

### Ein Modell ohne idiopathisch

Die Fehlerrate des Modelles liegt bei 0.28. Also wurde das Modell durch das Löschen von allen idiopathischen Fällen um ca. 12 % besser, obwohl der Datensatz nur noch 494 Patienten hat. Es ist klar, dass das Modell besser wird, wenn man die Anzahl Output Variablen reduziert. Doch in diesem Fall ist die Verbesserung dennoch bemerkenswert, denn der Datensatz wurde mehr als halbiert.

Die genutzten Laborwerte sind:

1. Location
2. c-reactive
3. mch
4. rdw
5. abn\_od\_haze
6. hemoglobin
7. proteinase-3\_antibodies
8. rv
9. abn\_os\_haze
10. abn\_od\_cells
11. neutrophil\_%

### Ein Modell ob idiopathisch oder nicht

Hier liegt die Fehlerquote bei 0.26. Das Modell hat eine gute Fehlerrate und ist auch leicht erklärt. Dieses wird in folgendem [Bild](http://data/DTVIZ-idio.PNG) dargestellt.

Die genutzten Laborwerte sind:

1. location
2. rv
3. abn\_os\_haze
4. Lysozyme-plasma
5. hemoglobin

### Löschen von unwichtigen Laborwerten

Es hat sich gezeigt, dass es einige Laborwerte gibt, die fast nie vom Modell genutzt werden. Diese sind: rheumatoid\_factor, hbs\_ag, hcv\_ab, beta-2-microglobulin, anti-ena\_screen, wbc, gender, cme, pars planitis, brao.

Im Weiteren sind gewisse kategorielle Diagnosen zu selten, um eine gute Vorhersage zu gewährleisten. Diese sind: not uveitis, scleritis, unknown

Der Decision Tree hat eine Fehlerrate von 0.4. Also blieb die Qualität des Modelles in etwa gleich. Das heisst, dass das Löschen keinen positiven Effekt hat, aber auch keinen negativen. Die oben genannten Laborwerte sind in der Tat nicht in den Modellen drin. Dies ist ein Hinweis, dass gesamthaft betrachtet, die obigen Laborwerte nicht wichtig sind.

Das Modell mit den gelöschten Laborwerten ergibt auch für das Random Forest Modell keine Verbesserung.

# Sarkoidose

Eine Studie untersucht den Zusammenhang zwischen einem erhöhten ACE- Wert und geringem Lymphozyten- Wert im Blutplasma. Cotta et al sagen in ihrer Studie, dass Uveitis anhand von erhöhten ACE Werten und niedrigen Lymphozyten- Werte bei vorhandener Sarkoidose diagnostiziert werden kann. (Source 3)

## Ergebnis Sarkoidose

Der Datensatz wurde nach Patienten gefiltert die einen erhöhten ACE -Wert haben, was die Anzahl Patienten von 1075 auf 111 reduzierte. Nach der Filterung nach einem niedrigen Lymphozyten- Wert blieben nur noch 13 Patienten übrig. Wobei keiner dieser Patienten ein Granulom aufwies.

Es hat sich gezeigt, dass eine Kombination aus diesen Werten im Datensatz gar nicht vorkommt. Also kann das Resultat der Studie weder bestätigt noch widerlegt werden. Es ist aber ein Hinweis, dass ein Krankheitsbild, wie in der Studie angenommen, sehr selten ist. Was die Frage aufwirft, ob es sinnvoll ist, eine Uveitis- Diagnose auf so seltene Gegebenheiten durchzuführen.

# Zusammenfassung

Die erstellten Modelle und deren Fehlerrate lassen darauf schliessen, dass eine Vorhersage mit diesen Daten problematisch ist. Die Verteilung der Output Variable ist sehr unterschiedlich, wodurch vor allem die Vorhersage des Random Forstes gelitten hat. Ein weiterer wichtiger Punkt ist die unterschiedliche Gewichtung der Laborwerte. Die Location wird vom Modell fast immer als wichtigstes Kriterium für die Klasse gesehen. Und weil die Verteilung der kategoriellen Diagnose in diesen Locations sehr unterschiedlich ist. Kann es sinnvoll sein die kategoriellen Diagnose in den Locations separat durchzuführen. Allerdings ist es wichtig, dass man für alle Locations möglichst gleich viele Daten hat. Ansonsten kann es sein, dass nur Vorhersagen in gewissen Locations funktionieren.

Es hat sich gezeigt, dass es einige unwichtige Labortest gibt. Diese verbessern weder das Modell, noch verschlechtern sie es. Auch gibt es kategorielle Diagnosen, die sehr selten sind und daher das Modell eher verschlechtern.

Die Verwendung von Idiopathischen Fällen ist eines der grössten Probleme des Datensatzes. Ein Modell mit weniger Daten und ohne Idiopathische Fälle schneidet besser ab als eines mit doppelt so vielen Daten und mit idiopathischen Fällen. Und eine Vorhersage ob idiopathisch oder nicht schneidet, mit 25% Fehlerrate, sogar fast am beste ab. Man sollte sich ernsthaft die Überlegung machen, diese überhaupt im Datensatz zu belassen.

## Bewertung der Methoden

Die grösste Frage ist, ob eine Fehlerrate von 40% als gut bezeichnet werden kann. Ich persönlich bin der Meinung das dieser Wert zu hoch ist. Denn es gibt Modelle die einfach alles als idiopathisch klassifizieren und eine Fehlerquote von 50% haben. Die Studie über die Sarkoidose hat allerdings auch einen Fehlerwert von 40%. Menschen sind sehr unterschiedlich und es kann gut sein, dass bei gewissen Entdeckungen eine Fehlerquote von 40% als gut eingestuft wird. Ich habe versucht mit diversen Methoden die Fehlerquote zu verbessern, was mit in den meisten Fällen nicht gelungen ist. Ich glaube aber, dass Uveitis Diagnosen mit Hilfe des Decision Trees oder sogar des Random Forsts einen sehr geringen Wert erreichen könnten, wenn der Datensatz symmetrischer gestaltet wird.

Die pars Platinis ist in 99% der Fälle eine intermediate idiopathische Uveitis. Dies verfälscht das Ergebnis, da man denkt pars Platinis seit leicht vorherzusagen. Dies ist zwar der Fall, aber es verfälscht das Ergebnis. Denn eine intermediate Uveitis kann auch systemisch sein.

## Persönliches Fazit

Die Berechnungen wurden für den Kompetenz- Erwerb durchgeführt. Es haben sich während der Challenge viele Umstände ergeben, die den Fokus der Fragestellung in eine andere Richtung gelenkt haben. Wenn ich alles, was ich über den Datensatz und die Modellberechnung gelernt habe, schon am Anfang gewusst habe, hätte ich einige Entscheidungen anders getroffen. Da ich zu Beginn keine Erfahrung in der Verwendung des Decision Tree Packages hatte, ist der Code auch nicht sehr sauber gehalten. Dass heisst ich hätte einige Codezeilen weglassen können oder die Variabel Namen besser sein könnten.

## Quellen

Pic1: <https://www.meraner.eu/gesundheit/rheuma-und-augenentzuendung-uveitis.2087?page=>

Pic 2: <https://zhengtianyu.wordpress.com/2013/12/13/decision-trees-and-entropy-algorithm/>

Pic 3: Erstellt von Alexandre Rau für die Präsentation: [Riccard\pp\Uveitis-data.pptx](http://riccard/pp/Uveitis-data.pptx)

Pic 4: Erstellt von Riccard Nef im Berechnung Notebook: [uveitsCalculation.Rmd](http://uveitscalculation.rmd)

Source1 : <https://towardsdatascience.com/imbalanced-data-oversampling-using-genetic-crossover-operators-bb581150a9a8>

Source 2 : <https://www.upgrad.com/blog/gini-index-for-decision-trees/>

Source 3: Cotte P, Pradat P, Kodjikian L, et alDiagnostic value of lymphopaenia and elevated serum ACE in patients with uveitisBritish Journal of Ophthalmology Published Online First: 11 September 2020. doi: 10.1136/bjophthalmol-2020-316563