Universität Zürich[UZH]

**Soziologisches Institut**

**Term Paper**

# Klassifikation von Übergewicht anhand von Soziodemografischen Merkmalen

**Fortgeschrittene Statistik**


| | |
|---|---|
| Verfasser*in: | **Samuel Rauh** |
| Matrikel-Nr.: | 20-734-067 |
| Email: | samuel.rauh@uzh.ch |


| | |
|---|---|
| Modulname: | Fortgeschrittene Statistik |
| Modulkürzel: | Modulkuerzel, z.B. 240-MAa240-010a |
| Semesterangabe: | HS22 |
| Betreuer*in/Dozent*in: | Dr. Marco Giesselmann |
| Abgabedatum: | 15.01.2023 |

# 1 Einleitung

Die Gesundheitskosten sind ein Thema, welches viele Menschen in der Schweiz belasten. Dieses Jahr werden die Krankenkassenprämien wieder steigen, welche bereits einen grossen Kostenpunkt für viele Menschen am Ende des Monats ausmachen. Im Sorgenbarometer, welches Jährlich von der Credit Suisse erhoben wird stehen die Gesundheitskosten im Jahr 2022 auf Platz 6, noch vor dem Ukraine Krieg (Credit Suisse, 2022).
Übergewicht ist ein relevantes Gesundheitsrisiko, es kann zum Beispiel ein (Mit-)Auslöser für Diabetes oder kardiovaskuläre Erkrankungen sein (Abdullah et al., 2010; Ernst et al., 1997).
Das Ziel dieser Arbeit ist die Klassifikation von Übergewicht mittels statistischer Methoden. Ich gehe von einem realistischen Szenario aus, in welchem eine kantonale oder nationale Behörde eine Klassifikation von Übergewichtigen Personen braucht, um eine Präventionskampagne zu lancieren und so möglichst zukünftige Gesundheitsrisiken zu mindern und die Kosten des Gesundheitssystem zu senken.
Beruhend auf dem kreierten Szenario, gehe ich davon aus, dass solch eine Behörde Zugriff auf persönliche Daten hat, die zum Beispiel aus Dokumenten wie der Steuererklärung stammen.
Diese Informationen sollen dabei helfen, eine bessere Aussage darüber zu treffen, ob eine Person übergewichtig ist oder nicht.
Verschiedenste soziodemografische Merkmale können einen Einfluss auf die Chance haben, dass eine Person übergewichtig ist oder nicht.
Schlechte Ernährung und fehlende körperliche Aktivität zum Beispiel sind Determinanten von Übergewicht, welche beide durch fehlende Bildung gefördert werden. Wer sich den negativen Auswirkungen von ungesunder Ernährung und mangelnder Bewegung nicht bewusst ist, wird sich auch weniger bemühen, diesen Dingen nachzugehen (Kriwy & Jungbauer-Gans, 2020).
Das Einkommen kann zum Beispiel einen Einfluss auf die Wahrscheinlichkeit von Übergewicht haben über die Ernährung oder die Möglichkeit, physische Aktivitäten auszuüben. Nährstoffreiche, kalorienarme und daher gesündere Nahrung ist oft teurer als nährstoffarme, kalorienreiche Ernährung. Wenn man bei den Essenseinkäufen mehr auf den Preis schauen muss, greift man demnach auch häufiger zu Ernährung, welche Übergewicht fördert (Kriwy & Jungbauer-Gans, 2020). Weiter greifen ärmere Familien eher zu Nahrungsmittel, bei welchen sie wissen, dass ihre Kinder sie mögen, um möglichst effizient einkaufen zu können. Dabei handelt es sich auch eher um kalorienreiche und nährstoffarme Nahrungsmittel, so wird auch schon früh der Geschmack der Kinder an solche Nahrungsmittel angepasst und es gibt weniger Möglichkeiten, diesen auch an gesünderes zu gewöhnen (Daniel, 2016). Auch sportliche Aktivitäten wie ein Abonnement in einem Fitnessstudio oder sogar ein Personal-Trainer können ein hoher Kostenpunkt sein, der sich ärmere Personen nicht leisten können.
Für die Bildung und das Einkommen gibt es jedoch auch den umgekehrten Ansatz, dass Personen mit Übergewicht weniger erfolgreich in Ihrer Karriere sind und somit auch die Chance sinkt, jemanden mit gleichem oder höherem Sozioökonomischem Status zu heiraten (von Hippel & Lynch, 2014).
Durch die körperliche Attraktivität hat auch der Zivilstand eine Auswirkung auf die Wahrscheinlichkeit von Übergewicht. Ein gesundes Gewicht gilt als äusserlich attraktiver und somit ist auch die Chance höher, dass man einen Partner oder eine Partnerin findet (Kriwy & Jungbauer-Gans, 2020).
Die Unterschiede zwischen den Geschlechtern könnten auch ein Resultat von Schönheitsnormen sein. Das Gewicht könnte bei Schönheitsidealen für Frauen eine grössere Rolle spielen als bei Männern. Somit würden sich Frauen mehr Gedanken über Ihr Gewicht machen, sich demnach gesünder ernähren und sich mehr bewegen (Kanter & Caballero, 2012).

Es gibt eine Vielzahl von Studien, sie sich mit Determinanten von Übergewicht befassen. Eine Studie aus China zum Beispiel hat gezeigt, dass ein tiefer sozioökonomischer Status mit der Wahrscheinlichkeit von Übergewicht zusammenhängt. Bei Frauen korrelieren ein tieferes Einkommen, wie schlechtere Bildung mit der Chance auf Übergewicht, bei Männern ist es nur das Bildungsniveau (Zhang et al., 2017).
Smith et al. (2012) fanden bei einer Untersuchung von US-Militär Personal heraus, dass grundsätzlich Männer häufiger von Übergewicht betroffen sind als Frauen, zusätzlich korrelierte Übergewicht mit höherem Alter, afro-amerikanischer oder lateinamerikanischer Ethnizität und verheiratetem Zivilstand.

# 2 Methode

## 2.1 Daten

Die Daten, welche als Grundlage dieser Klassifikation dienen, stammen aus dem Schweizer Haushaltspanel (Tillmann, o. J.). Das Schweizer Haushaltspanel (SHP) ist eine umfangreiche, jährlich durchgeführte Sozialwissenschaftliche Studie. Bei den Daten hier handelt es sich um die Welle 17, welche zwischen September 2015 und Februar 2016 erhoben wurden. Als Erhebungsmethode wird hier meist CATI (Computer assisted Telephone Interview) verwendet, in seltenen Fällen finden auch Interviews vor Ort statt (Tillmann et al., 2022). Bei der Datengrundlage für diese Arbeit handelt es sich um den Teildatensatz analytical_file im STATA Format.
Wie bereits erwähnt beruht die Auswahl der Variablen darauf, welche Informationen eine Behörde zur Verfügung hat. Grundsätzlich können nun alle diese Variablen miteinbezogen werden, da das einzige Ziel dieser Klassifikation darin besteht, eine möglichst hohe Genauigkeit zu erreichen. In Tabelle 1 sind die Ausgewählten Variablen und die Anzahl gültiger Einträge ersichtlich.

Tabelle 1: Übersicht der benutzten Variablen

| **Variablenübersicht** | **Overall (N=13‘950)** |
|---|---|
| **BMI** | |
| Gültig | 9'226 |
| **Alter** | |
| Gültig | 13'930 |
| **Geschlecht** | |
| Gültig | 13'950 |
| **Internationale Standard-Klassifikation von Bildung ISCED 1997** | |
| Gültig | 13'880 |
| **Jährliches Totaleinkommen, netto** | |
| Gültig | 8'328 |
| **Gemeinde Typologie** | |
| Gültig | 13'948 |
| **Anzahl Kinder** | |
| Gültig | 9'257 |
| **Zivilstand** | |
| Gültig | 13'945 |
| **ISCO-Klassifikation: Hauptberuf** | |
| Gültig | 5'950 |
| **Ersparnisse** | |
| Gültig | 8'199 |

Um die Klassifikationsalgorithmen zu trainieren, müssen aus dem Datensatz alle nicht auswertbaren Werte (NA) entfernt werden. Somit bleibt eine Stichprobe von 4'933 Personen.

## 2.2 Operationalisierung

Die abhängige Variable des Übergewichtes ist eine kreierte Variable, die von der Körpergrösse und dem Körpergewicht abgeleitet wird. Zuerst wird der Body Mass Index (BMI) wie folgt berechnet: $Gewicht\ in\ kg \Big/ körpergrösse\ in\ m^2$ . Für die Ableitung, ob es sich bei dem Gewicht für die entsprechende Körpergrösse um Übergewicht oder Untergewicht handelt, werden verschiedene Schwellwerte gesetzt. Normallerweise geht man bei einem BMI von 25.0 und höher von Übergewicht aus, während man ab 30.0 von Adipositas spricht (Apovian, 2016). Ich habe mich dafür entschieden mich mit der Klassifikation von Übergewicht zu befassen, da hier eine genug grosse Stichprobe besteht, wie in Abbildung 1 ersichtlich.

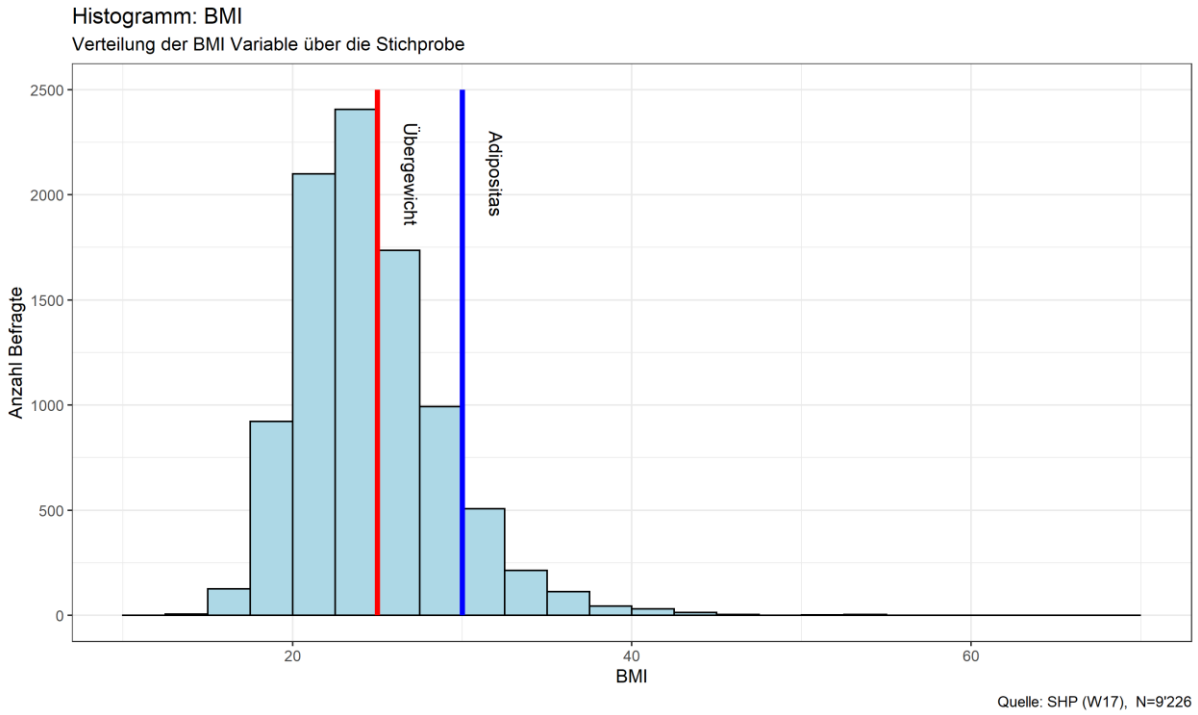


Abbildung 1: Univariate Verteilung BMI

Die determinierende Variable Einkommen wir hier operationalisiert als das Jährliche Nettoeinkommen in Franken, wie auch die Ersparnisse. Der Bildungsgrad ist operationalisiert in 10 ordinalen Kategorien nach ISCED Standard. Dazu kommt der Zivilstand, die Anzahl der Kinder und das Geschlecht in binärer Form. Der Wohnort ist im SHP nicht mit der Postleitzahl angegeben, dies würde bei solch einer kleinen Stichprobe jedoch auch wenig Sinn ergeben. Stattdessen sind die Gemeinden in neun Kategorien eingeteilt (z.B. Zentren, periphere städtische Gemeinden oder wohlhabende Gemeinden).
Das gleiche gilt für die Berufe, diese sind in zehn Kategorien wie «Techniker und assoziierte Fachleute» oder «Handwerker und verwandte Berufe» eingeteilt.

## 2.3 Logistische Regression vs. Decision-Tree

Wie bereits in der Einleitung erwähnt, wird bei dieser Arbeit von einem realistischen Szenario ausgegangen. Eine Behörde will mit einer Präventionskampagne möglichst viele Übergewichtige Personen erreichen und dabei möglichst wenige nicht übergewichtige als solche kategorisieren. Ich setze für diese Arbeit das Ziel, 70% aller übergewichtigen Personen als solche zu erkennen und dabei möglichste wenige falsche positive zu erhalten.
Um dieses Ziel zu erreichen, werde ich zwei verschiedene Methoden benutzen und diese vergleichen: die logistische Regression und die Decision-Tree Klassifikation. In Abschnitt 3 werden die Resultate der Klassifikation der beiden Methoden verglichen und dargelegt, welche Methode sich für solch eine Aufgabe besser eignet.
Da es bei dieser Arbeit nicht darum geht, einen Zusammenhang zu erkennen, sondern eine Voraussage zu treffen, muss der Datensatz in einen Satz Trainingsdaten und einen Satz Testdaten geteilt werden. Ich habe mich hier für eine Teilung von 60/40 entschieden, das heisst 60% der Merkmalsträger werden verwendet um die Modelle zu «trainieren» und 40% werden verwendet, um sie zu testen. Besonders für das Decision-Tree verfahren ist diese Unterteilung sehr wichtig, da es sonst zur Überanpassung kommen kann, wobei sich die Modelle zu stark an einen bestimmten Datensatz angepasst sind und sich so die Klassifizierungs-Leistung auf Basis der Grundgesamtheit verschlechtert.

Die logistische Regression ist in der Statistik eine verbreitete Methode zur Klassifizierung von binären Variablen und ist somit auch für diese Aufgabe geeignet. Die logistische Regression wird durch folgende Funktion beschrieben:

$$\frac{1}{1 + e^{-(-0.76+0.02 * Alter - 0.90 * Geschlecht\ (Frau) - 0.000001 * Einkommen)}}$$

Dabei handelt es sich nicht um die komplette Regressionsgleichung, sondern nur um eine Ausgewählte Sammlung von Variablen zur Veranschaulichung. Um die oben beschriebene Grenze von 70% der Übergewichtigen mit der logistischen Regression zu erreichen, bin ich wie folgt vorgegangen. Die logistische Regression ist ein Verfahren zur Klassifikation von binären Variablen, doch wie der Name scho sagt, handelt es sich dabei um eine Regression. Fügt man nun also die Werte der unabhängigen Variablen in die Gleichung ein, ergibt sich daraus einen Wert zwischen 0 und 1. Im Normalfall werden nun Resultate von über 0.5 als 1 und von unter 0.5 als 0 klassifiziert. Bei dieser Aufgabe geht es jedoch darum, einen bestimmten Anteil an Übergewichtigen Personen zu finden. Damit das sicher gestellt werden kann, habe ich eine Methode entwickelt, in welcher der Schwellwert von 0.5 aus, iterativ verschoben wird, bis zwischen 69.9% und 70.1% der Übergewichtigen als solche klassifiziert werden.

Das Decision-Tree Verfahren funktioniert so, indem, wie der Name sagt, ein Entscheidungsbaum erstellt wird. Dabei wird bei jedem Knoten (Nodes) ein Schwellenwert (bei metrischen Variablen) oder eine Einteilung von Kategorien (bei nicht-metrischen Variablen) für eine der unabhängigen Variablen bestimmt, welche den Pfad dann in zwei Teile teilt. Dies kann dann beliebige male geschehen, je nachdem wie die gewünschten Parameter gewählt werden. Zum Schluss landet jeder Pfad in einem „Bucket“, der eine der Kategorien der abhängigen Variable repräsentiert

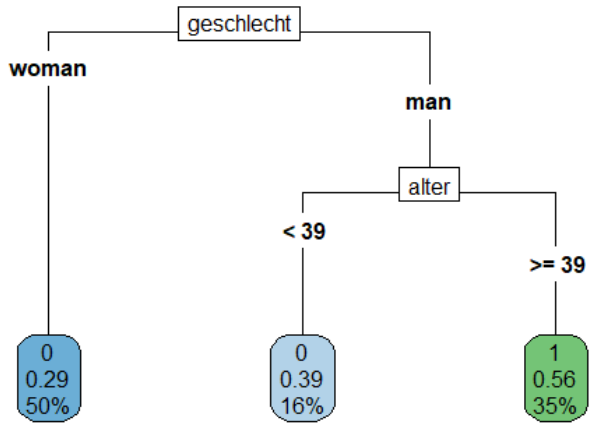


Abbildung 2: Beispiel Decision-Tree

Abbildung 2 zeigt eine vereinfachte Methode des Entscheidungsbaumes mit lediglich 2 Nodes und 3 Buckets. Dabei handelt es sich nicht um den realen Baum und dient lediglich der Veranschaulichung und Erklärung der Decision-Tree Klassifikation.

Um die Genauigkeit der Klassifikation des Decision-Trees zu verbessern, gibt es verschiedene Parameter, die man anpassen kann. Ein wichtiger ist der „complexity parameter“ (cp) (der Name kann je nach verwendetem Package variieren), vereinfacht gesagt gibt er an, wie viel sich die Vorhersagequalität mit jedem Split verbessern muss. Je kleiner er ist, desto mehr Teilungen gibt es und desto höher ist die Vorhersagegenauigkeit für die jeweilige Stichprobe. Hier muss jedoch beachtet werden, dass ein zu kleiner cp Wert zu sehr hohem Rechenbedarf führt und zusätzlich kann dies zu „Over Fitting“ führen. Dabei passt sich der Algorithmus zu sehr der Trainings-Stichprobe an und so verschlechtert sich die Klassifikationsfähigkeit für die generelle Grundgesamtheit.

Es gibt auch viele weitere Parameter, z. B. kann die maximale Tiefe festgelegt werden, heisst wie viele Male höchstens geteilt wird. Man kann auch bestimmen, wie viele Observierungen in einem Node existieren müssen, damit er sich weiter aufteilen kann. Mit Hilfe einer „Grid-Search“, bei welcher alle möglichen Kombinationen für verschiedene Ausprägungen von gewählten Parametern ausprobiert werden, wurden die besten Parameter bestimmt.

Da der Decision-Tree eine Klassifikation in Kategorien vornimmt und nicht einen Wert zwischen 0 und 1 retourniert wie die logistische Regression, muss hier auf eine andere Methode zurückgegriffen werden, damit der gewünschte Anteil von 70% der übergewichtigen Personen gefunden werden kann. Hier wird der „Prior-Parameter“ angepasst. Dieser gibt an, wie eine Variable in der Grundgesamtheit verteilt sein sollte. Wenn man diese Manipuliert, verschiebt sich die Vorhersagewahrscheinlichkeit. Dieser Parameter wird dann wieder in einem Iterativen Verfahren so verändert, bis zwischen 69,9% und 70,1% der Übergewichtigen gefunden wurden.

# 3 Resultate

Die Resultate der beiden Klassifikationsalgorithmen sind in den Tabellen 2 (logistische Regression) und 3 (Decision-Tree) ersichtlich. Die Kreuztabellen zeigen die eigentliche Verteilung im Test-Datensatz und die Resultate der beiden Klassifikationsalgorithmen. In den oberen linken und unteren rechten Zellen sind Anzahl von korrekt Klassifizierten Übergewichtsvariable (Ja-Nein) dargestellt. Oben recht zeigt die übergewichtigen Personen, welche nicht als solche klassifiziert wurden, während unten links nicht übergewichtige Personen, welche fälschlicherweise als solche klassifiziert wurden, dargestellt sind.
Das Ziel der Algorithmen war, 70% aller Übergewichtigen als solche zu erkennen und dabei möglichst wenige falsche-Positive zu erreichen. So ist auch ersichtlich, dass die beiden Klassifikationen eine sehr ähnliche Anzahl an korrekten Positiven aufweisen, auf welche die Methode abzielt.
Ein anderes Bild zeigt sich jedoch, wenn man auch die Personen miteinbezieht, welche nicht übergewichtig sind. Da die Algorithmen nicht perfekt sind, werden auch Personen, welche nicht übergewichtig sind, als solche klassifiziert. Besonders, da wir den Schwellenwert zur Klassifikation als übergewichtig nach unten setzen mussten, da ansonsten nicht genug der übergewichtigen Personen als solche klassifiziert werden.
Die Logistische Regression klassifiziert 542 Personen fälschlicherweise als übergewichtig, während das Decision-Tree verfahren nur 520 Personen in dieser Kategorie hat. In solch einem Fall eignet sich der Precision-Score als Evaluationsmethode. Er wird wie folgt berechnet:

$$Precision - Score = \frac{Korrekte\ Positive}{Korrekte\ Positive + Falsche\ Positive}$$

Die logistische Regression erzielt einen Precision-Score von 0.504, das Decision-Tree-Verfahren einen Precision-Score von 0.516. Dabei handelt es sich nicht um einen sehr grossen Unterschied. Auf 100 als übergewichtig klassifizierte Personen liegt der Decision-Tree Algorithmus bei 52 Personen und die logistische Regression bei 50 Personen richtig, es trennen sie also weniger als 2 richtige Klassifikationen.
Dabei handelt es sich jedoch um eine starke Verbesserung der Klassifikation im Gegensatz zu einer zufälligen Vorhersage. Ohne jegliche verbesserte Klassifikation müssten 70% aller Personen als übergewichtig eingestuft werden, um auch 70% aller übergewichtigen Personen zu erreichen. Da 41.5% der Stichprobe übergewichtig sind, hätte eine Klassifikation, ohne jegliches statistische Modell einen Precision-Score von 0.415, es gäbe also mehr falsche Positive als korrekte Positive.
Auch zu beachten ist, dass sich die allgemeine Performance der Algorithmen durch das Verschieben der Schwellwerte verschlechter.

Tabelle 2: Kreuztabelle: Klassifikationsresultat Logistische Regression

| | **Übergewicht** | **Wirklichkeit** | | **Total** |
|---|---|---|---|---|
| | | **Nein** | **Ja** | |
| **Logistische Regression** | **Nein** | **581** (70.17%) | 247 (29.83%) | 828 (43.12%) |
| | **Ja** | 542 (49.63%) | **550** (50.37%) | 1092 (56.88%) |
| | Total | 1123 (58.49%) | 797 (41.51%) | 1920 (100.00%) |

Tabelle 3: Kreuztabelle: Klassifikationsresultat Decision-Tree

| | Übergewicht | Wirklichkeit | | Total |
|---|---|---|---|---|
| | | Nein | Ja | |
| **Decision-Tree** | **Nein** | **603** (71.36%) | 242 (28.64%) | 845 (44.01%) |
| | **Ja** | 520 (48.37%) | **555** (51.63%) | 1075 (55.99%) |
| | Total | 1123 (58.49%) | 797 (41.51%) | 1920 (100.00%) |

Neben rein quantitativen Analysen der Resultate lassen sie die Klassifikationen auch qualitativ untersuchen. So können Lücken in den Algorithmen erkannt werden, welche sich zum Beispiel dadurch zeigen, dass eine bestimmte Gruppe von Personen häufig falsch klassifiziert wird.
Abbildung 3 zeigt, welcher Anteil der Personen je nach BMI richtig klassifiziert wurden. Dabei ist ersichtlich, dass bei beiden Methoden ein sehr ähnliches Muster Auftritt. Bei tiefem BMI ist die Fehlerrate sehr tief und steigt bis zum Schwellwert von 25 BMI-Punkten an. Dabei handelt es sich auch um ein Resultat, welches man erwarten würde, da Personen am nächsten beim Schwellwert die grössten Chancen haben, falsch klassifiziert zu werden. Dass dies besonders Personen gerade unter dem Schwellwert betrifft kommt davon, dass die Schwellwerte der Algorithmen zur Klassifizierung von Übergewicht heruntergesetzt wurden, um eine höhere Abdeckung übergewichtiger Personen zu erreichen.

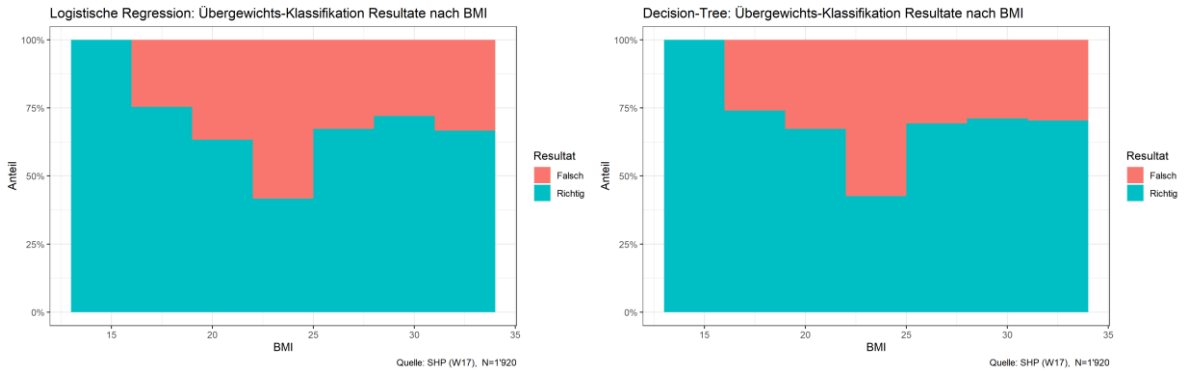


Abbildung 3: Resultate Klassifikation nach BMI

Weiter lässt sich in der Visualisierung des Decision-Tree und in der Regressionstabelle erkennen, welche Variablen für die Algorithmen in Betracht gezogen werden, beziehungsweise als Signifikant gekennzeichnet werden.
Wie man in Abbildung 4 sehen kann, entscheidet der Decision-Tree anhand der Variablen Geschlecht, Alter, Wohnort, Ersparnisse, Beruf und Bildung.

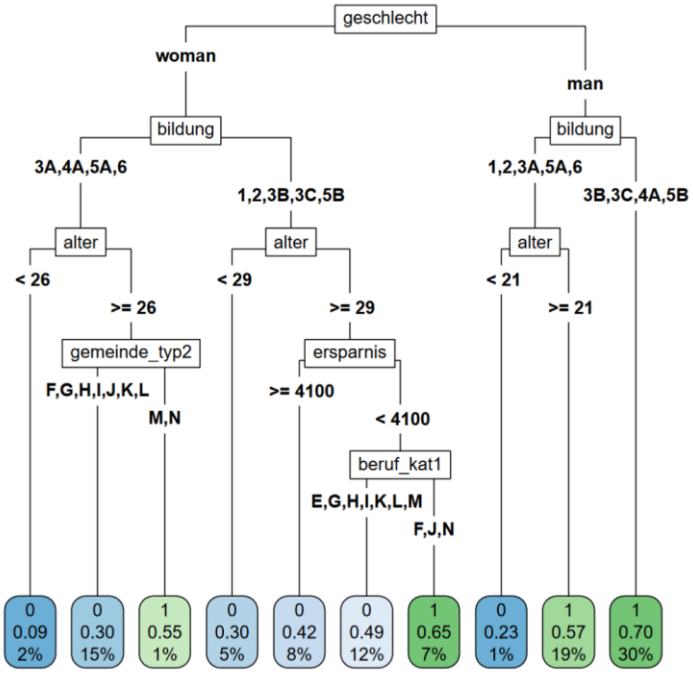


Abbildung 4: Decision-Tree Visualisierung (Variablenwerte Rekodiert zu Veranschaulichungszwecken, Zuweisung im Anhang ersichtlich)

Es ist erstaunlich, dass der Decision-Tree aus all den Variablen nur von diesen Gebrauch macht, schliesslich werden bei der logistischen Regression auch der Zivilstand und die Anzahl Kinder als signifikante Determinanten von Übergewicht eingeordnet. Ich gehe davon aus, dass der Grund für dieses Ergebnis die kleine Stichprobe ist. Bei solch kleinen Stichproben kann es schnell zu Over-Fitting kommen, und so kann man mit einem vereinfachten Modell bessere Resultate erzielen.

Tabelle 4: Logistische Regression

| | Übergewicht |
|---|---|
| Alter | 0.013*** |
| Geschlecht: Frau (ref: Mann) | -0.976*** |
| Bildung: Zweite Stufe Tertiärer Abschluss (ref: Primarstufe) | -0.794* |
| gemeinde_typ2Industrial and tertiary sector communes | 0.237* |
| gemeinde_typ2Rural commuter communes | 0.352** |
| Anzahl Kinder | -0.090** |
| Zivilstand: Verheiratet (ref: nie verheiratet) | 0.268** |
| Zivilstand: Verwitwet | 0.848*** |
| Constant | -1.917 |
| Observations | 3,013 |
| Log Likelihood | -1,886.313 |
| Akaike Inf. Crit. | 3,846.626 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

Die Resultate beider Klassifikationsmethoden widerspiegeln auch die theoretischen Herleitungen und die Ergebnisse der Literaturrecherche. Durch die verbesserte Klassifikation wird bestätigt, dass die soziodemografischen Merkmale einen Einfluss auf die Wahrscheinlichkeit von Übergewicht haben. Tabelle 4 und Abbildung 4 zeigen auch, dass sich die Resultate weitgehend mit der Literatur decken, zum Beispiel sind Männer und ältere Personen häufiger von Übergewicht betroffen.

# 4 Diskussion

Wie in der Theorie und in der Literatur beschrieben, haben soziodemografische Merkmale einen Einfluss auf die Wahrscheinlichkeit, dass eine Person übergewichtig ist. Dies zeigt sich in der verbesserten Vorhersage von Übergewicht mit Hilfe der statistischen Modellen.
Zusammengefasst zeigen die Resultate, dass sich das Decision-Tree Verfahren besser als Klassifikationsalgorithmus eignet im Gegensatz zur logistischen Regression. Mit ein paar Ausnahmen werden vom Decision-Tree Variablen verwendet, welche auch in der logistischen Regression als signifikante Determinanten bestimmt werden und auch über die Verteilung des BMIs betrachtet, verhalten sich die beiden Klassifikationen recht ähnlich.
Bei einer solchen Aufgabe gibt es viele verschiedene Limitationen. Eine erste ist die Operationalisierung von Übergewicht mittels des BMI. Es ist ein viel diskutiertes Thema, ob es Sinn macht, sich bei der Klassifizierung von Übergewicht nur auf die Körpergrösse und das Gewicht zu beziehen, besonders da sich dieses Verhältnis mit dem Alter oder dem Geschlecht grundsätzlich verändern kann (Nuttall, 2015). Dies ist jedoch eine Limitation, die einen Inhaltlichen Aspekt betrifft, welcher hier nicht im Fokus steht. Bei einer wirklichen Anwendung wäre es jedoch ein Faktor, welcher nicht ausser Acht gelassen werden darf.
Es gibt auch viele methodische Limitationen. Ein grosser Punkt, der bereits angesprochen wurde, ist die Stichprobengrösse. Viele Algorithmen können erst ab einer genug grossen Stichprobe ihr volles Potential entwickeln, der Decision-Tree Algorithmus gehört in gewissem Masse auch dazu. Dies hat sich beim Trainieren der Algorithmen gezeigt, so musste ich das Verhältnis des Train-Test-Splits zugunsten des Test-Datensatzes verändern, was beim Decision-Tree für bessere Resultate gesorgt hat. Man kann also davon ausgehen, dass sich die Leistung des Decision-Trees mit einer grösseren Stichprobe verbessern würde, dann könnten wirklich auch alle Determinanten Variablen verwendet werden, ohne dass es direkt zu Over Fitting kommt. Auch die logistische Regression würde sich mit einer grösseren Stuckprobe etwas verbessern, jedoch nicht im selben Mass. Es könnte sogar sein, dass mit einer kleineren Auswahl an Variablen ein besseres Resultat der Klassifikation hätte erzielt werden können. Grundsätzlich führen mehr Variablen zu einer grösseren Erklärungskraft und somit auch zu einer verbesserten Klassifikation, jedoch würden eine kleinere Auswahl an Variablen zu einer grösseren Stichprob führen, da so weniger Merkmalsträger mit NAs wegfallen würden. In einem echten Szenario gäbe es auch weitere Möglichkeiten NAs zu behandeln. Zum Beispiel könnte man sie mit Median Werten der jeweiligen Variable ausfüllen oder sogar mit einem Algorithmus anhand der bestehenden Variablen eine best-mögliche Schätzung abgeben. Dies würde jedoch den Rahmen dieses Projektes sprengen.
Eine weitere Limitation betrifft die Optimierung des Decision-Trees. Man kann nie sicher sein, ob man die perfekten Werte für alle Parameter gefunden hat, man ist somit eigentlich nie am Ende der Optimierung angelangt.
In den Resultaten habe ich untersucht, bei welchen BMI-Werten die Klassifikationsalgorithmen besser bzw. schlechtere Leistungen gezeigt haben. Dabei habe ich jedoch nur die Verteilung des BMIs in Betracht gezogen. Das gleiche könnte man auch für alle anderen Variablen durchführen, zum Beispiel

mit dem Alter. So könnten weitere wichtige Einsichten erlangt werden, um die Klassifikation zu verbessern.
Eine weitere Limitation für den Decision Tree ist der Prior-Faktor. Er wird verändert, um das gewünschte Verhältnis von als übergewichtig klassifizierten Personen zu erreichen. Es ist jedoch nicht klar, welche Auswirkung seine Veränderung auf die Performance des Algorithmus hat. Dies ist ein grundsätzliches Problem, wenn man tiefer in die Materie von Maschinellem Lernen geht, da irgendwann nicht mehr ersichtlich ist, was genau in der "Black Box" passiert. Es wird auch argumentiert, dass bei Algorithmen mit starken Auswirkungen für die Öffentlichkeit, grossen Wert auf die Interpretierbarkeit gelegt werden soll, um potentiellen Schaden an der Gesellschaft zu vermeiden (Rudin, 2019).

# 5 Literaturverzeichnis


Abdullah, A., Peeters, A., de Courten, M., & Stoelwinder, J. (2010). The magnitude of association between overweight and obesity and the risk of diabetes: A meta-analysis of prospective cohort studies. *Diabetes Research and Clinical Practice*, *89*(3), 309–319. https://doi.org/10.1016/j.diabres.2010.04.012

Apovian, C. M. (2016). Obesity: Definition, Comorbidities, Causes, and Burden. *THE AMERICAN JOURNAL OF MANAGED CARE*, *22*(7).

Credit Suisse. (2022). *Sorgenbarometer 2022: Grosse Rochade bei den Top-Sorgen – Spitzenreiter Umwelt und Altersvorsorge*. https://www.credit-suisse.com/about-us-news/de/articles/media-releases/credit-suisse-sorgenbarometer-2022--grosse-rochade-bei-den-top-s-202211.html

Daniel, C. (2016). Economic constraints on taste formation and the true cost of healthy eating. *Social Science & Medicine*, *148*, 34–41. https://doi.org/10.1016/j.socscimed.2015.11.025

Ernst, N. D., Obarzanek, E., Clark, M. B., Briefel, R. R., Brown, C. D., & Donato, K. (1997). Cardiovascular Health Risks Related to Overweight. *Journal of the American Dietetic Association*, *97*(7), S47–S51. https://doi.org/10.1016/S0002-8223(97)00729-3

Kanter, R., & Caballero, B. (2012). Global Gender Disparities in Obesity: A Review. *Advances in Nutrition*, *3*(4), 491–498. https://doi.org/10.3945/an.112.002063

Kriwy, P., & Jungbauer-Gans, M. (Hrsg.). (2020). *Handbuch Gesundheitssoziologie*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-06392-4

Nuttall, F. Q. (2015). Body Mass Index: Obesity, BMI, and Health A Critical Review. *Nutrition Today*, *50*(3), 117–128. https://doi.org/10.1097/NT.0000000000000092

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Smith, T. J., Marriott, B. P., Dotson, L., Bathalon, G. P., Funderburk, L., White, A., Hadden, L., & Young, A. J. (2012). Overweight and Obesity in Military Personnel: Sociodemographic Predictors. *Obesity*, *20*(7), 1534–1538. https://doi.org/10.1038/oby.2012.25

Tillmann, R. (o. J.). *Living in Switzerland Waves 1-22 + Beta version wave 23 + Covid 19 data* (11.0) [Data set]. FORS. https://doi.org/10.48573/PW95-E774

Tillmann, R., Voorpostel, M., Antal, E., Dasoki, N., Klaas, H., Kuhn, U., Lebert, F., Monsch, G.-A., & Ryser, V.-A. (2022). The Swiss Household Panel (SHP). *Jahrbücher Für Nationalökonomie Und Statistik*, *242*(3), 403–420. https://doi.org/10.1515/jbnst-2021-0039

von Hippel, P. T., & Lynch, J. L. (2014). Why are educated adults slim—Causation or selection? *Social Science & Medicine*, *105*, 131–139. https://doi.org/10.1016/j.socscimed.2014.01.004

Zhang, H., Xu, H., Song, F., Xu, W., Pallard-Borg, S., & Qi, X. (2017). Relation of socioeconomic status to overweight and obesity: A large population-based study of Chinese adults. *Annals of Human Biology*, *44*(6), 495–501. https://doi.org/10.1080/03014460.2017.1328072

# Anhang: R Code

```
library(tidyverse)
library(readr)
library(ggplot2)
library (stargazer)
library(summarytools)
library(haven)
library(visreg)
library(magrittr)
library(labelled)
library(dplyr)
library(rpart)
library(rpart.plot)
library(caTools)
library(caret)
library(table1)
library(crosstable)
library(officer)
library(flextable)

setwd("Path")
analytical_file = read_dta('Data/analyticalfile_full.dta')
```

Preprocessing:

```
#varlist <- look_for(analytical_file)

# attributes(analytical_file$isced17)
# freq(analytical_file$isced17)

data <- analytical_file %>%
  select(gewicht = p17c46,
         körpergrösse = p17c45,
         alter = age17,
         geschlecht = sex17,
         bildung = isced17,
         einkommen = i17ptotn,
         gemeinde_typ2 = com2_17,
         kinder = ownkid17,
         zivilstand = civsta17,
         beruf_kat1 = is1maj17,
         ersparnis = p17i165
         )

data$BMI <- data$gewicht / ((data$körpergrösse/100)^2)
data$geschlecht <- as_factor(data$geschlecht)
# data$gemeinde_typ1 <- as_factor(data$gemeinde_typ1)
data$gemeinde_typ2 <- as_factor(data$gemeinde_typ2)
data$beruf_kat1 <- as_factor(data$beruf_kat1)
# data$beruf_kat2 <- as_factor(data$beruf_kat2)
data$zivilstand <- as_factor(data$zivilstand)
```

```
data$bildung <- as_factor(data$bildung)

summary_table <- table1(data=data, ~.)
```

Univariate Analyse BMI

```
bmi_hist <- ggplot(data, aes(x=BMI))+
  geom_histogram(fill='lightblue', color='black', breaks=seq(10, 70, 2.5))
+
  geom_segment(aes(x = 25, y = 2500, xend = 25, yend = 0), color="red", si
ze=1.5)+
  annotate("text", x=27, y=2100, label="Übergewicht", angle=270)+
  geom_segment(aes(x = 30, y = 2500, xend = 30, yend = 0), color="blue", s
ize=1.5)+
  annotate("text", x=32, y=2100, label="Adipositas", angle=270)+
  theme_bw()+
  labs(title='Histogramm: BMI',
       subtitle = "Verteilung der BMI Variable über die Stichprobe",
       y = 'Anzahl Befragte',
       x =  'BMI', caption = "Quelle: SHP (W17),  N=9'226")
ggsave('Exports/bmi_hist.png', plot = bmi_hist, width = 10, height = 6)

data$übergewicht[data$BMI>=25] <- 1
data$übergewicht[data$BMI<25] <- 0

data$adipositas[data$BMI>=30] <- 1
data$adipositas[data$BMI<30] <- 0
```

Train-Test-Split

```
set.seed(111)

data_noNA <- na.omit(data)


ind <- sample(2, nrow(data_noNA),
              replace = T,
              prob = c(0.6, 0.4))
train <- data_noNA[ind==1,] %>% subset(select = -c(BMI, adipositas, körper
grösse, gewicht))
test <- data_noNA[ind==2,] %>% subset(select = -c(BMI, adipositas, körperg
rösse, gewicht))
test_bmi <- data_noNA[ind==2,] %>% subset(select = c(BMI))

# determinanten_train <- subset(train, select = -c(BMI, adipositas, körper
grösse, gewicht, übergewicht))
result_train <- select(train, übergewicht)
#
# determinanten_test <- subset(test, select = -c(BMI, adipositas, körpergr
össe, gewicht, übergewicht))
result_test <- select(test, übergewicht)
```

Logistische Regression Vereinfachte logistische Reression zur veranschaulichung

```
logr <- glm(data = train, family = binomial(link = "logit"),
            übergewicht ~ .)
# summary(logr)


log_simple <- glm(data = train, family = binomial(link = "logit"),
            übergewicht ~ alter + geschlecht + einkommen)
summary(log_simple)
```

Funktion um Treshhold für logistische Regression festzulegen

```
logr <- glm(data = train, family = binomial(link = "logit"),
            übergewicht ~ .)

test_tresh <- test

set_treshhold_log <- function(model, train, result, percentage) {
  # train$prediction_log <- predict(model, newdata = train, type="response
")
  test_tresh$prediction_log <- predict(model, newdata = test, type="respon
se")

  split_at = 0.5

  while (TRUE) {
    # train$pred_log_bin[train$prediction_log >= split_at] <- 1
    # train$pred_log_bin[train$prediction_log < split_at] <- 0

    test_tresh$pred_log_bin[test_tresh$prediction_log >= split_at] <- 1
    test_tresh$pred_log_bin[test_tresh$prediction_log < split_at] <- 0

    # c_table <- ctable(train$pred_log_bin, result, prop = "t")
    c_table <- ctable(test_tresh$pred_log_bin, result_test, prop = "t")

    cross_table <- c_table$cross_table

    found_positive <- cross_table[5]
    total_positive <- cross_table[6]

    found_perc <- found_positive/total_positive


    if (between(found_perc, percentage-0.01, percentage+0.01)) {
      print(cross_table)
      print(paste0("Treshhold found at splitting point:", split_at))
      return(split_at)
    } else if (found_perc < percentage) {
      split_at <- split_at - (percentage - found_perc)/10
    } else if (found_perc > percentage) {
      split_at <- split_at + (percentage + found_perc)/10
    }
```

```
  }
}

threshold_log <- set_treshhold_log(logr, train, result_train, 0.7)
```

Decision Tree Grid-Search: beste kombination von Parametern finden

```
set.seed(111)

grid_search <- list(minsplit = c(30, 33, 35, 40, 50, 70),
           maxdepth = c(3, 4, 5, 6),
           cp = c(0.01, 0.003, 0.001)) %>%
  cross_d()

mod <- function(...) {
  rpart(data=train, übergewicht ~ ., control = rpart.control(...), method
= "class")
}

grid_search <- grid_search %>% mutate(fit = pmap(grid_search, mod))

compute_accuracy <- function(fit, test_features, test_labels) {
  predicted <- predict(fit, test_features, type = "class")
  mean(predicted == test_labels)
}

test_features <- test %>% select(-übergewicht)
test_labels   <- test$übergewicht
grid_search <- grid_search %>%
  mutate(test_accuracy = map_dbl(fit, compute_accuracy,
                                 test_features, test_labels))

grid_search <- grid_search %>%
  arrange(desc(test_accuracy), desc(minsplit), maxdepth, desc(cp))
grid_search[1:5,]
minsplit_1 <- grid_search$minsplit[1]
maxdepth_1 <- grid_search$maxdepth[1]
cp_1 <- grid_search$cp[1]
```

Prior Parameter finden um 70% der Übergewichtigen zu finden

```
set.seed(111)

test_prior <- test

set_treshhold_tree <- function(train, result, percentage) {
  prior_perc = 0.4
  while (TRUE) {
    tree <- rpart(data=train, übergewicht ~ .,
            control = rpart.control(minsplit = minsplit_1,
                                    maxdepth = maxdepth_1,
                                    cp = cp_1),
```

```
            method = "class",
            parms = list(prior = c(prior_perc, 1-prior_perc)))

    # train$prediction_tree <- predict(tree, newdata = train, type="class
")

    prediction_tree <- predict(tree, test_features, type='class')
    test_prior$prediction_tree <- prediction_tree



    # c_table <- ctable(train$prediction_tree, result, prop = "t")
    c_table <- ctable(test_prior$prediction_tree, result_test, prop = "t")

    cross_table <- c_table$cross_table

    found_positive <- cross_table[5]
    total_positive <- cross_table[6]

    found_perc <- found_positive/total_positive

    if (between(found_perc, percentage-0.01, percentage+0.01)) {
      print(cross_table)
      print(paste0("Prior percentage fount at:", prior_perc))
      return(prior_perc)
    } else if (found_perc < percentage) {
      prior_perc <- prior_perc - (percentage - found_perc)/10
    } else if (found_perc > percentage) {

      prior_perc <- prior_perc + (percentage + found_perc)/10
    }
  }
}

prior_tree <- set_treshhold_tree(train, result_train, 0.7)
```

Logistische Regression Resutate

```
test$log_pred <- predict(logr, newdata = test, type="response")

test$log_pred_bin[test$log_pred >= threshold_log] <- 1
test$log_pred_bin[test$log_pred < threshold_log] <- 0

test$log_result[test$log_pred_bin==result_test$übergewicht] <- "Richtig"
test$log_result[test$log_pred_bin!=result_test$übergewicht] <- "Falsch"

log_table <- ctable(test$log_pred_bin, result_test)
print(log_table, file = 'Exports/log_table.html')

test$log_result <- relevel(as.factor(test$log_result), ref = "Falsch")
```

```
log_plot <- cbind(test, test_bmi) %>%
  ggplot(aes(x=BMI, fill=log_result))+
  geom_histogram(position = "fill", binwidth = 5, breaks = seq(13, 35, 3))
+
  labs(title = "Logistische Regression: Übergewichts-Klassifikation Result
ate nach BMI",
       y="Anteil", x="BMI", fill = "Resultat",
       caption = "Quelle: SHP (W17),  N=1'920")+
  scale_y_continuous(labels = scales::percent_format(scale = 100, accuracy
 = 1))+
  theme_bw()
log_plot

ggsave("Exports/log_results_vis.png", plot = log_plot, width = 8, height =
 5)

log_crosstable <- crosstable(data = test, c(log_pred_bin), by=übergewicht,
 total = "both") %>%
  as_flextable()
log_crosstable
save_as_docx(log_crosstable, path = "Exports/log_crosstable.docx")
```

Decision-Tree Resultate

```
set.seed(111)

tree <- rpart(data=train, übergewicht ~ .,
              control = rpart.control(minsplit = minsplit_1,
                                      maxdepth = maxdepth_1,
                                      cp = cp_1),
              method = "class",
              parms = list(prior = c(prior_tree, 1-prior_tree)))
test$tree_pred <- predict(tree, test_features, type='class')

test$tree_result[test$tree_pred==result_test$übergewicht] <- "Richtig"
test$tree_result[test$tree_pred!=result_test$übergewicht] <- "Falsch"


tree_table <- ctable(test$tree_pred, result_test)

test$tree_result <- relevel(as.factor(test$tree_result), ref = "Falsch")

tree_plot <- cbind(test, test_bmi) %>%
  ggplot()+
  geom_histogram(aes(x=BMI, fill=tree_result), position = "fill", breaks =
 seq(13, 35, 3))+
  labs(title = "Decision-Tree: Übergewichts-Klassifikation Resultate nach
BMI",
       x="BMI", y="Anteil", fill="Resultat",
       caption = "Quelle: SHP (W17),  N=1'920")+
  scale_y_continuous(labels = scales::percent_format(scale = 100, accuracy
 = 1))+
```

```
  theme_bw()
tree_plot

ggsave("Exports/tree_results_vis.png", plot = tree_plot, width = 8, height
 = 5)

tree_crosstable <- crosstable(data = test, c(tree_pred), by=übergewicht, t
otal = "both") %>%
  as_flextable()
tree_crosstable
save_as_docx(tree_crosstable, path = "Exports/tree_crosstable.docx")

stargazer(logr, type = "html",
          out = "Exports/logr_table.doc")

summary(logr)
```

Visualisierung Decision-Tree

```
set.seed(111)

# attributes(train$bildung)
# attributes(train$beruf_kat1)
# attributes(train$gemeinde_typ2)

# Ukodierung, da ansonsten die Variablenbeschreibungen keinen Platz haben
auf der Visualisierung
recode_bildung <- c("0: Not completed primary (compulsory) education" = "0
",
                    "1: Primary or first stage of basic education" = "1",
                    "2: Lower secondary or Second stage of basic education
" = "2",
                    "3A: Upper secondary education (preparation for tertia
ry education)" = "3A",
                    "3B: Upper secondary education (preparation for furthe
r prof. education)" = "3B",
                    "3C: Upper secondary education (entrance into the labo
r market)" = "3C",
                    "4A: Post-secondary education non tertiary (preparatio
n for an institution for higher education)" = "4A",
                    "5A: First stage of tertiary education (general educat
ion)" = "5A",
                    "5B: First stage of tertiary education (professional e
ducation)" = "5B",
                    "6: Second stage of tertiary education" = "6" )
recode_beruf <- c("no corresponding ISCO-value" = "A",
                  "other error, unplausible value" = "B",
                  "inapplicable" = "C",
                  "no answer" = "D",
                  "Armed forces" = "E",
                  "Legislators, senior officials, managers" = "F",
                  "Professionals" = "G",
```

```
                    "Technicians and associate professionals" = "H",
                    "Clercs" = "I",
                    "Service workers, market sales workers" = "J",
                    "Skilled agricultural and fishery workers" = "K",
                    "Craft and related trades workers" = "L",
                    "Plant and machine operater assemblers" = "M",
                    "Elementary occupations" = "N")

recode_gemeinde <- c("other error" = "A",
                    "filter error" = "B",
                    "inapplicable" = "C",
                    "no answer" = "D",
                    "does not know" = "E",
                    "Centres" = "F",
                    "Suburban communes" = "G",
                    "Wealthy communes" = "H",
                    "Peripheral urban communes" = "I",

                    "Tourist communes" = "J",
                    "Industrial and tertiary sector communes" = "K",
                    "Rural commuter communes" = "L",
                    "Mixed agricultural communes" = "M",
                    "Peripheral agricultural communes" = "N")

plot_data <- train %>%
  mutate(bildung = recode(bildung, !!!recode_bildung),
         beruf_kat1 = recode(beruf_kat1, !!!recode_beruf),
         gemeinde_typ2 = recode(gemeinde_typ2, !!! recode_gemeinde))

tree_for_plot <- rpart(data=plot_data, übergewicht ~ .,
            control = rpart.control(minsplit = minsplit_1,
                                    maxdepth = maxdepth_1,
                                    cp = cp_1),
            method = "class",
            parms = list(prior = c(prior_tree, 1-prior_tree)))

rpart.plot(tree_for_plot, type = 5)
```

Universität Zürich[UZH]

## Soziologisches Institut

# Selbständigkeitserklärung

| | |
|---|---|
| Titel der Arbeit*: | **Klassifikation von Übergewicht anhand von Soziodemografischen Merkmalen** |
| Modulname: | **Fortgeschrittene Statistik** |
| Betreuer*in/Dozent*in: | **Dr. Marco Giesselmann** |

* nur Haupttitel - ohne Untertitel

Ich erkläre ausdrücklich, dass es sich bei der von mir eingereichten schriftlichen Arbeit um eine von mir selbst und ohne unerlaubte Beihilfe sowie in eigenen Worten verfasste Originalarbeit handelt. Sofern es sich dabei um eine Arbeit von mehreren Verfasserinnen oder Verfassern handelt, bestätige ich, dass die entsprechenden Teile der Arbeit korrekt und klar gekennzeichnet und der jeweiligen Autorin oder dem jeweiligen Autor eindeutig zuzuordnen sind.

Ich bestätige überdies, dass die Arbeit als Ganzes oder in Teilen weder bereits einmal zur Abgeltung anderer Studienleistungen an der Universität Zürich oder an einer anderen Universität oder Ausbildungseinrichtung eingereicht worden ist, noch inskünftig durch mein Zutun als Abgeltung einer weiteren Studienleistung eingereicht werden wird.

**Verwendung von Quellen**
Ich erkläre ausdrücklich, dass ich sämtliche in der oben genannten Arbeit enthaltenen Bezüge auf fremde Quellen (einschliesslich Tabellen, Grafiken u. Ä.) als solche kenntlich gemacht habe. Insbesondere bestätige ich, dass ich ausnahmslos und nach bestem Wissen sowohl bei wörtlich übernommenen Aussagen (Zitaten) als auch bei in eigenen Worten wiedergegebenen Aussagen anderer Autorinnen oder Autoren (Paraphrasen) die Urheberschaft angegeben habe.

**Sanktionen**
Ich nehme zur Kenntnis, dass Arbeiten, welche die Grundsätze der Selbstständigkeitserklärung verletzen – insbesondere solche, die Zitate oder Paraphrasen ohne Herkunftsangaben enthalten – als Plagiat betrachtet werden und die entsprechenden rechtlichen und disziplinarischen Konsequenzen nach sich ziehen können (gemäss §§ 7ff der Disziplinarordnung der Universität Zürich sowie § 39der Rahmenverordnung für das Studium in den Bachelor- und Master-Studiengängen der Philosophischen Fakultät der Universität Zürich).

**Ich bestätige mit meiner Unterschrift die Richtigkeit der Angaben.**

| | |
|---|---|
| Verfasser*in: | Samuel Rauh |
| Matrikelnummer: | 20-734-067 |

Zürich, 15.01.2023 Unterschrift: