Inhalt

I. Datenanalyse Stroke Prediction

1. Datenquelle

Kaggle: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

2. Basisinformationen der Daten *(Kennzahlen)*

Wir betrachten hier kurz die wesentlichen Grundinformationen zu den Daten.

2.1 Spalten

1) id: unique identifier - ID	(fortlaufender Identifier)
2) gender – Geschlecht	(3 Kategorien, Text)
3) age - Alter	(kontinuierliche Werte)
4) hypertension - Bluthochdruck	(binäre Werte)
5) heart_disease – Herzkrankheit	(binäre Werte)
6) ever_married – verheiratet (früher oder heute)	(2 Kategorien, Text)
7) work_type – Beruflicher Status	(5 Kategorien, Text)
8) Residence_type - Wohnstatus (urban oder städtisch)	(2 Kategorien, Text)
9) avg_glucose_level – durchschnittl. Blutzuckerwert	(kontinuierliche Werte)
10) bmi – Body Mass Index	(kontinuierliche Werte)
11) smoking_status – Status Raucher	(4 Kategorien, Text)
12) stroke – Infarkt	(binär) - **Zielspalte**

2.2 Fehlende Werte

```
 #   Column             Non-Null Count
---  ------             --------------
 0   id                 5110 non-null
 1   gender             5110 non-null
 2   age                5110 non-null
 3   hypertension       5110 non-null
 4   heart_disease      5110 non-null
 5   ever_married       5110 non-null
 6   work_type          5110 non-null
 7   Residence_type     5110 non-null
 8   avg_glucose_level  5110 non-null
 9   bmi                4909 non-null
 10  smoking_status     5110 non-null
 11  stroke             5110 non-null
```

Beim BMI liegen 201 fehlende Werte vor (NaN). Diese werden wir später auffüllen.

2.3 Wertebereiche und statistische Grunddaten

Die binären Spalten haben die Einträge 0 oder 1 mit folgenden Häufigkeiten:

hypertension	
0	4612
1	498

heart_disease	
0	4834
1	276

Die **kategorialen Spalten** haben folgende Kategorien mit entsprechender Häufigkeit:

gender	
Female	2994
Male	2115
Other	1

ever_married	
Yes	3353
No	1757

Residence_type	
Urban	2596
Rural	2514

work_type	
Private	2925
Self-employed	819
children	687
Govt_job	657
Never_worked	22

smoking_status	
never smoked	1892
Unknown	1544
formerly smoked	885
smokes	789

Die **kontinuierlichen Spalten** haben folgende Wertebereiche und statistische Grunddaten:

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.000000	5110.000000	5110.000000	5110.000000	5110.000000	4909.000000	5110.000000
mean	36517.829354	43.226614	0.097456	0.054012	106.147677	28.893237	0.048728
std	21161.721625	22.612647	0.296607	0.226063	45.283560	7.854067	0.215320
min	67.000000	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	17741.250000	25.000000	0.000000	0.000000	77.245000	23.500000	0.000000
50%	36932.000000	45.000000	0.000000	0.000000	91.885000	28.100000	0.000000
75%	54682.000000	61.000000	0.000000	0.000000	114.090000	33.100000	0.000000
max	72940.000000	82.000000	1.000000	1.000000	271.740000	97.600000	1.000000

Da wir später auf Plots zu den Daten blicken, verzichten wir hier auf weitere Erläuterungen. Folgendes Stellen wir aber bereits fest:

stroke	
0	4861
1	249

Es liegen sehr wenige Datenzeilen vor, die eine 1 bei stroke haben. Wir haben also **unbalancierte Klassen y=0 und y=1**. Das erschwert die Prognose, wie wir sehen werden.

Ein kurzer Blick auf die Unterschiede zwischen Männern und Frauen bei den Durchschnittswerten und der Standardabweichung:

Männer

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	2115.000000	2115.000000	2115.000000	2115.000000	2115.00000	2011.000000	2115.000000
mean	36562.541371	42.483385	0.104965	0.077069	109.08852	28.647936	0.051064
std	21146.470229	23.484066	0.306580	0.266763	47.43484	7.464493	0.220180

Frauen

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	2994.000000	2994.000000	2994.000000	2994.000000	2994.000000	2897.000000	2994.000000
mean	36479.685037	43.757395	0.092184	0.037742	104.057809	29.065758	0.047094
std	21176.443056	21.966561	0.289334	0.190604	43.590651	8.110783	0.211876

Man sieht:

- Frauen sind in den Daten leicht in der Überzahl. Der Anteil an strokes ist bei Männern und Frauen jedoch ähnlich (5,1% bei Männern, 4,7% bei Frauen).
- Die restlichen Daten zeigen im Durchschnitt keine auffälligen Unterschiede, außer dass die Männer deutlich häufiger an Herzkrankheit leiden.

3. Kodierung der kategorischen Spalten

Wir ziehen hier zugunsten der nachfolgenden Visualisierungen hier bereits einen Teil der Datenvorbereitung vor. Wir haben die kategorialen Spalten gender, heart_disease, ever_married, work_type und Residence_type, die wir für die Algorithmen codieren müssen.

Wir haben uns für den Ordinal Encoder entschieden, und dabei folgende Kategorien gewählt:

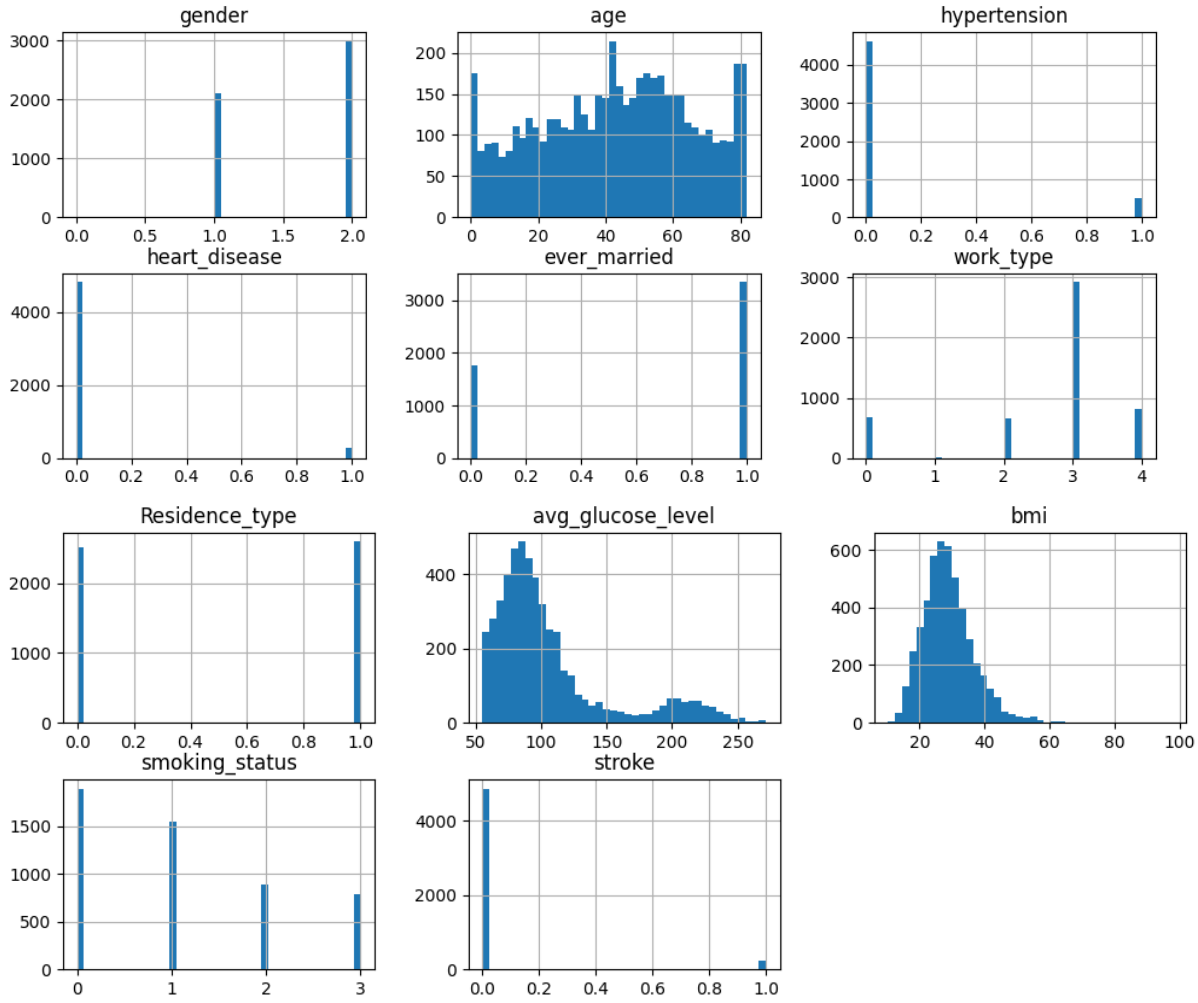
```
Spalte: gender
  Other -> 0
  Male -> 1
  Female -> 2
Spalte: ever_married
  No -> 0
  Yes -> 1
Spalte: work_type
  children -> 0
  Never_worked -> 1
  Govt_job -> 2
  Private -> 3
  Self-employed -> 4
Spalte: Residence_type
  Rural -> 0
  Urban -> 1
Spalte: smoking_status
  never smoked -> 0
  Unknown -> 1
  formerly smoked -> 2
  smokes -> 3
```

Dabei ist bewusst, dass höhere vergebene integer-Werte für eine Korrelation relevant sind. Nach online-Recherche bezüglich Risikofaktoren (siehe Quellenangabe am Ende der Datenanalyse) für Infarkte und eigene Überlegungen / Allgemeinwissen zu Gesundheit haben wir die Kategorien in den obigen Reihenfolgen kodiert. Höhere Werte vermuten wir als Risikofaktoren für strokes. **Achtung**: Es ist bekannt, dass Frauen häufiger strokes haben als Männer. Aber wir haben aufgrund der sehr niedrigen entstandenen Korrelation (siehe später bei der Korrelationsmatrix) entschieden, dass die Codierung so trotzdem ok ist.

4. Inhalte der Daten

4.1 Histogramme

Im folgenden die Histogramme für die Häufigkeiten aller Spalten. Darunter folgen unsere Beobachtungen.

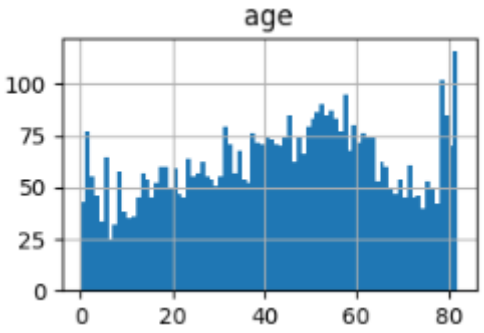


Beobachtungen zu Histogrammen

- ungleiche Klassenverteilungen bei: hypertension, heart_disease, und vor allem stroke!
    - Das ist ungünstig. Die Zielklasse y=1 (strokes) ist unterrepräsentiert. Auch zwei der vermuteten Risikofaktoren (hypertension, heart disease) treten selten auf.

- age ist ganz grob gleichverteilt. Allerdings gibt es einen Häufigkeitspeak alter Personen, und überraschenderweise auch ganz junger Personen
- Residence_type in etwa 50/50
- glucose level:
    - normale Blutzuckerwerte: Linkssteil/rechtsschief verteilt; diabetische Blutzuckerwerte (ab ca. 175 aufwärts) in etwa normalverteilt, die beiden Verteilungen überlappen sich (Grenze ca. bei 175).
- bmi: rechtsschief/linkssteil

Wenn wir nochmals auf das Alter schauen mit höherer Auflösung (mehr bins) sehen wir nochmals, dass im Alter und in jungen Jahren viele Daten vorliegen (hier noch nicht ersichtlich, ob mit stroke oder nicht).



4.2 Statistische Werte Gesunde Patienten und stroke-Patienten

Wir betrachten kurz die Durchschnitte / Standardabweichungen für stroke = 0 und stroke = 1:

Stroke = 1:

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
count	249.000000	249.000000	249.000000	249.000000	249.000000	249.000000	249.000000	249.000000	209.000000	249.000000	249.0
mean	1.566265	67.726908	0.265060	0.188755	0.883534	3.104418	0.542169	132.544739	30.471292	1.257028	1.0
std	0.496588	12.734166	0.442254	0.392102	0.321429	0.675866	0.499222	61.921056	6.329452	1.120826	0.0

Stroke = 0:

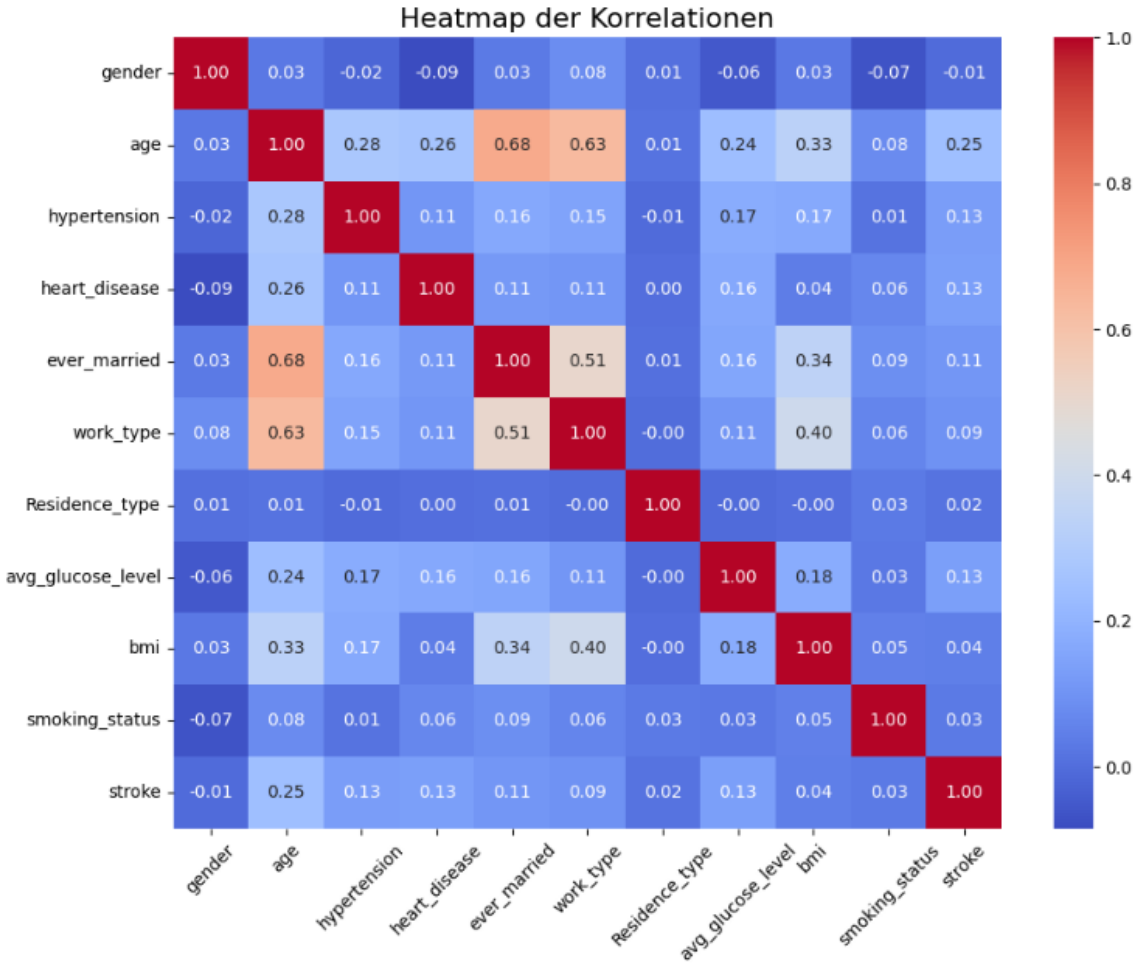
	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
count	4861.000000	4861.000000	4861.000000	4861.000000	4861.000000	4861.000000	4861.000000	4861.000000	4700.000000	4861.000000	4861.0
mean	1.586711	41.959679	0.088871	0.047110	0.644518	2.594939	0.506274	104.795513	28.823064	1.104300	0.0
std	0.492892	22.313775	0.284586	0.211895	0.478708	1.185773	0.500012	43.846069	7.908287	1.068837	0.0

Es fällt auf bei stroke = 1:

- höherer Altersdurchschnitt
- Risikofaktoren im Durchschnitt erhöht:
    - Herzkrankheit, Bluthochdruck, Blutzuckerwerte, Raucher (nur leicht erhöht)

Damit können wir bereits Korrelationen vermuten. Wir schauen uns direkt die Heatmap der Korrelationen an.

4.3 Korrelationen



Beobachtungen zu Korrelationen

Überlegungen nützliche Spalten:

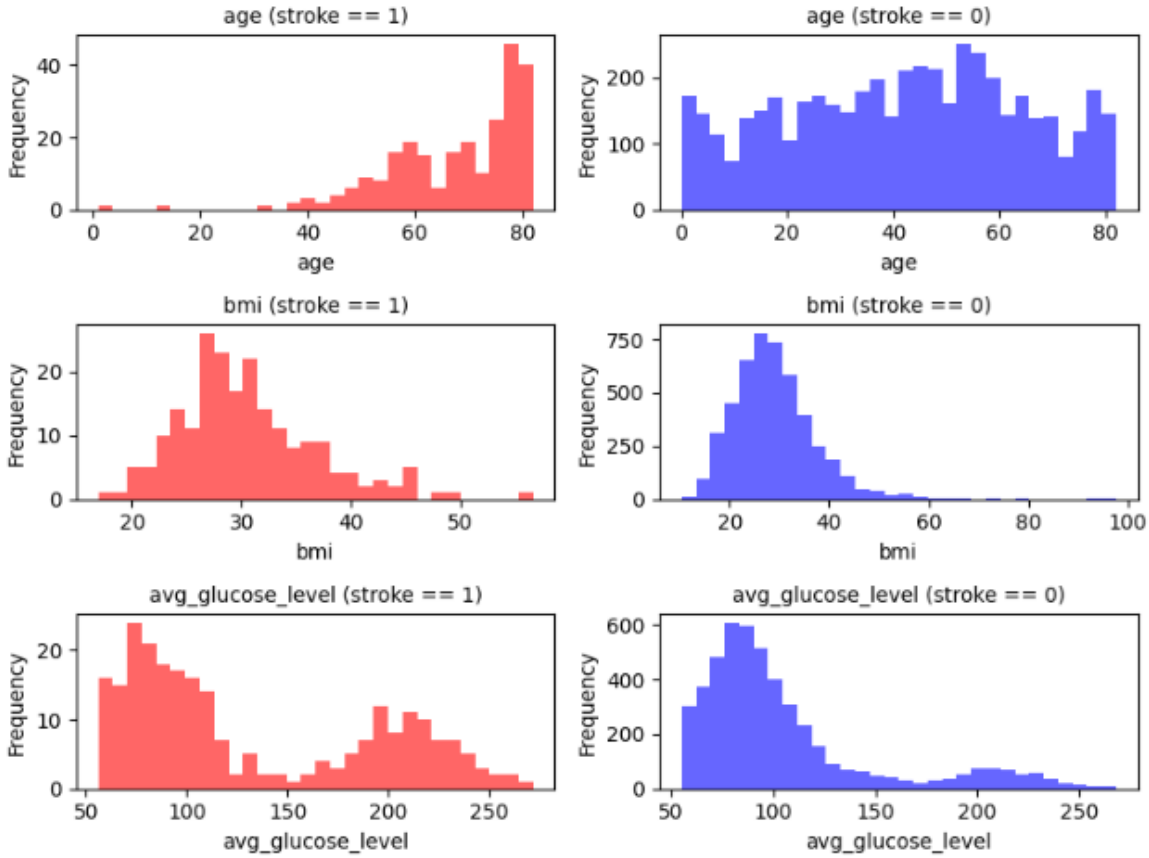
- insgesamt keine hohen Korrelationen zu stroke. Obige **Risikofaktoren** haben aber **höhere** Korrelation (trotzdem vergleichsweise niedrig):
  - am ehesten noch age (altersbedingtes Risiko)
  - ebenfalls noch hypertension, heart_disease, glucoselevel (als gesundheitliche Risikofaktoren)
- auf den ersten Blick überraschend: bmi und stroke: sehr niedrig korreliert. Interpretation: Alte Leute werden dünn, bekommen aber Infarkte -> korrelation von bmi/stroke sinkt.

Überlegung unnützliche Spalten:

- age und evermarried hohe Korrelation: **Ever_married lassen wir daher weg** (wer alt ist, hat ohnehin höheres Risiko, ever_married ist autokorreliertage und work_type hohe Korrelation: evt. work_type weglassen?
- ever_married und work_type sind selbst auch sehr korreliert! Wir **lassen daher später work_type auch weg**.

4.4 Visualisierung Verteilungen bei stroke = 0 und stroke = 1

Kontinuierliche Spalten



Beobachtungen kontinuierliche Spalten:

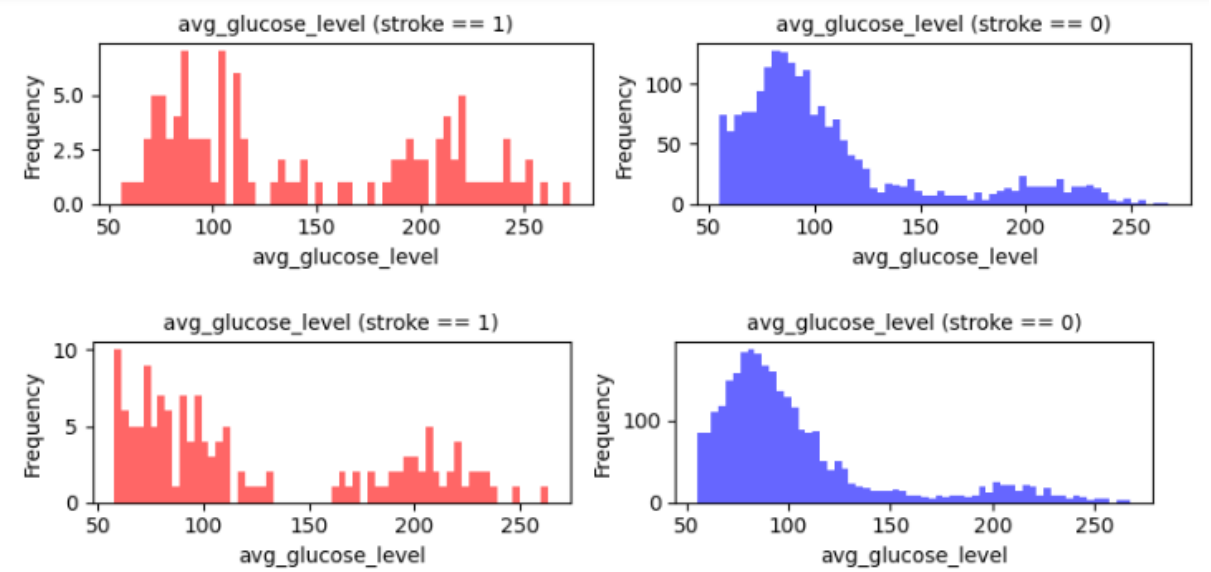
- age:
  - bei stroke =1 zunehmend ältere Patienten
  - ABER: in JEDEM Alter gibt es Patienten mit stroke (wenige) und auch ohne stroke (viele). Das erschwert die Prognose für die Algorithmen
- Glukosewerte und BMI: Selbe Problematik
- Bei Glukosewerten erkennbar: Ab Glukosewerten von ca. 150 steigt das Risiko für einen Herzinfarkt. Bei stroke = 0 ist das Höhenverhältnis zwischen rechtem und linkem Hügel deutlichgrößer als bei stroke = 1. Das bedeutet: Unter den stroke-Daten ist im Vergleich zu den gesunden der Anteil der Diabetiker (Blutzucker größer 150) erhöht!

Kurze getrennte Betrachtung Männer/Frauen:
Plottet man sich obige Histogramme getrennt für Männer und Frauen, stellt man keine bedeutenden Unterschiede fest, außer dass bei den Frauen zwei sehr junge Menschein einen Infarkt hatten (Alter 14 Jahre und 1 Jahr).

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
69768	2.0	1	0	0	0.0	0.0	1.0	70.37	NaN	1.0	1
49669	2.0	14	0	0	0.0	0.0	0.0	57.93	30.9	1.0	1

Da diese Personen keinerlei relevante Symptome für einen Stroke zeigen (Bluthochdruck, Herzkrankheit, etc. alles unauffällig), werden wir diese später entfernen.

Bei den Männern kann man noch erkennen, dass unter den Stroke-Patienten wiederum der Anteil der Diabetiker erhöht ist im Vergleich zu den gesunden Patienten. Im Plot unten (oben Männer, unten Frauen) erkennt man, dass bei stroke = 1 für Männer der rechte Hügel vergleichsweise steigt.

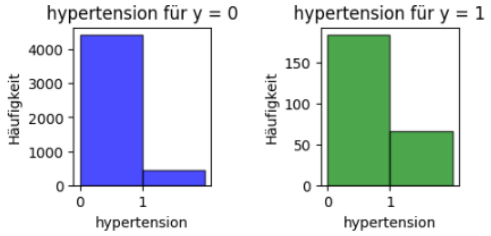


Fazit kontinuierliche Spalten

- Bei age lohnt sich evt. Zusatzspalte ab zB age > 60 (dort treten mehr strokes auf, bei Männern und bei Frauen)m, oder wir entfernen junge Patienten komplett
- bei glucose lohnt sich evt. Zusatzspalte bei Gluc > 150 (dort treten mehr strokes auf, bei Männern und bei Frauen)

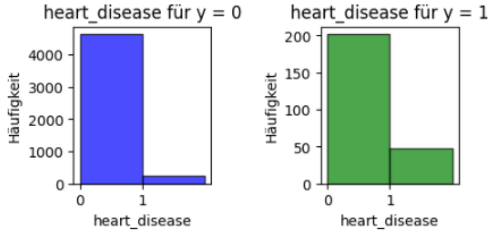
Kategoriale Spalten

Wir betrachten für die Kategorienspalten (2 Kategorien, oder mehr) Barplots und Kreuztabellen. Wir fangen mit folgenden dreien an, die wir als in den Daten abgebildete **Risikofaktoren** für strokes vermuten: **Bluthochdruck, Herzkrankheit, Rauchen**. Dabei steht rechts von der eigentlichen Kreuztabelle (0,1-Kreuzung für Attribut und stroke) die Summe des Attributs mit 0 bzw. 1 (Zeilensumme), der Anteil, den die Zeilensumme an den Gesamtdaten hat, und rechts der Prozentsatz für stroke = 0 bzw. stroke = 1 in der jeweiligen Zeile.
Am interessantesten ist hier zum einen die Kreuztabelle selbst, da wir ablesen können, ob ein Attribut die Klassen stroke = 0 und stroke = 1 sauber trennt, und die Spalte ganz rechts, die die bedingte Wahrscheinlichkeit für einen stroke angibt, gegeben den Fall, dass der Risikofaktor vorliegt!
Beispiel: Bei hypertension =1 und stroke = 1 stehen 66 Datensätze, diese machen 13.3 % der Zeilensumme 498 von hypertension = 1 aus. Eine Person mit Bluthochdruck bekommt also in unseren Daten mit 13,3% Wahrscheinlichkeit einen stroke (bei stochastischer Unabhängigkeit von anderen Attributen als vereinfachender Annahme). Wir können hier also schon die bedingten Wahrscheinlichkeiten ablesen, mit denen jemand einen stroke bekommt, wenn er eines der Risiken erfüllt (z.B. P(stroke =1|hypertension = 1) = 13,3%)!



Kreuztabelle

stroke	0	1	Zeilensumme	Anteil_an_Gesamt	healthy-%	stroke-%
hypertension						
0	4429	183	4612	90.3%	96.0%	4.0%
1	432	66	498	9.7%	86.7%	13.3%



Kreuztabelle

stroke	0	1	Zeilensumme	Anteil_an_Gesamt	healthy-%	stroke-%
heart_disease						
0	4632	202	4834	94.6%	95.8%	4.2%
1	229	47	276	5.4%	83.0%	17.0%

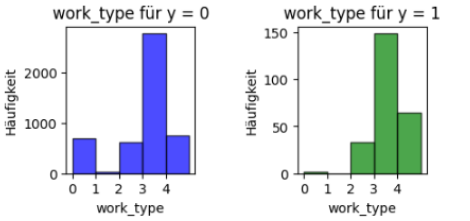
Kreuztabelle

stroke	0	1	Zeilensumme	Anteil_an_Gesamt	healthy-%	stroke-%
smoking_status						
0.0	1802	90	1892	37.0%	95.2%	4.8%
1.0	1497	47	1544	30.2%	97.0%	3.0%
2.0	815	70	885	17.3%	92.1%	7.9%
3.0	747	42	789	15.4%	94.7%	5.3%

Beobachtungen für diese Plots:

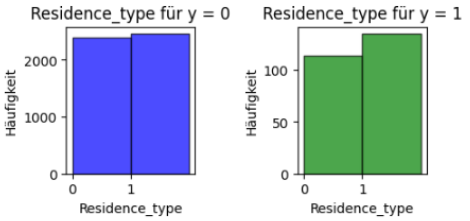
- Wir sehen überall:
  - Bei Vorliegen des jeweiligen Risikofaktors (bzw. beim Rauchen bei „aufsteigenden“ Kategorien) steigt auch die bedingte Wahrscheinlichkeit einen stroke zu erhalten (stroke-% ist bei Attribut = 0 immer kleiner als bei Attribut = 1 und höher!). Das deckt sich mit medizinischem Allgemeinwissen bzw. den Inhalten aus unseren Quellen
  - Trotzdem: Wir haben auch in den Risikogruppen (Risikofaktor ist erfüllt, also Attribut = 1) immer sehr viele Daten OHNE stroke. ´
- **Wir können hier schon ableiten:**
  - **Wir haben nie saubere „Nachbarschaften“ oder abtrennbare Bereiche, in denen wir eine Gruppe mit Risikofaktoren und strokes komplett vom Rest trennen können. Unsere Daten sind sehr durchmischt!**

Für die weiteren Spalten, die wir weniger als Risikofaktoren einschätzen, schauen wir uns ebenfalls die Plots/Kreuztabellen an.



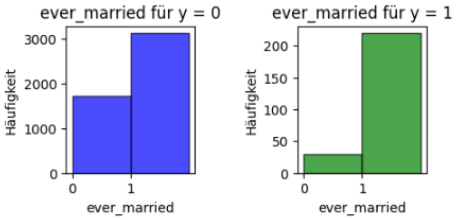
Crosstable

stroke	0	1	category_sum	share_category_of_total	healthy-%	stroke-%
work_type						
0.0	685	2	687	13.4%	99.7%	0.3%
1.0	22	0	22	0.4%	100.0%	0.0%
2.0	624	33	657	12.9%	95.0%	5.0%
3.0	2776	149	2925	57.2%	94.9%	5.1%
4.0	754	65	819	16.0%	92.1%	7.9%



Crosstable

stroke	0	1	category_sum	share_category_of_total	healthy-%	stroke-%
Residence_type						
0.0	2400	114	2514	49.2%	95.5%	4.5%
1.0	2461	135	2596	50.8%	94.8%	5.2%



Crosstable

stroke	0	1	category_sum	share_category_of_total	healthy-%	stroke-%
ever_married						
0.0	1728	29	1757	34.4%	98.3%	1.7%
1.0	3133	220	3353	65.6%	93.4%	6.6%

Hier sehen wir nichts, was uns besonders auffällt. Außer bei der Spalte ever_married, aber diese wollen wir weglassen (Autokorrelation zum Alter).

Fazit Kategoriale Spalten

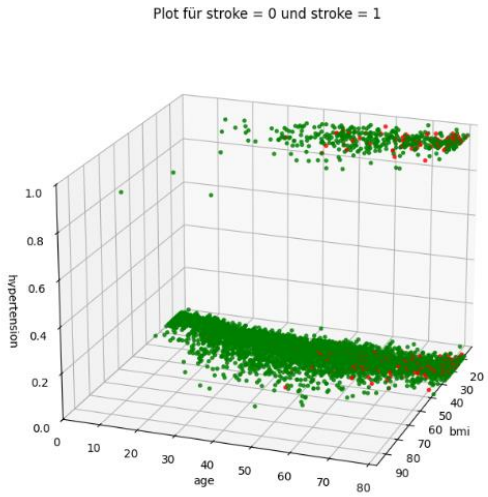
Wir sehen:

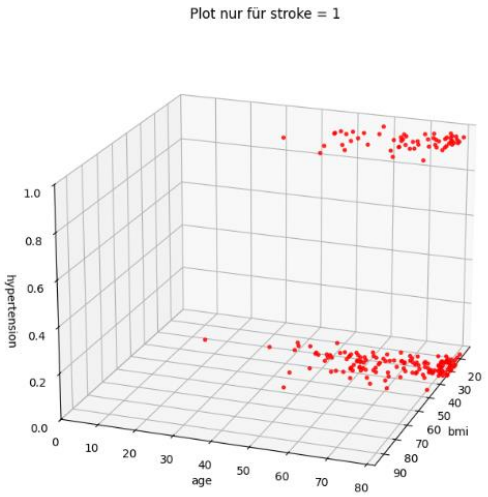
- Bei den Risikofaktoren steigt tatsächlich die bedingte Wahrscheinlichkeit auf einen Stroke, wenn das Risiko vorliegt (medizinisch interessant!)
- ABER: Wir haben trotzdem auch bei den Risiko-Attributen gemischte Ergebnisse bzgl. Stroke
- Insgesamt haben wir auch hier leider keine Spalte, bei der wir davon ausgehen können, dass sie große Datenteilbereiche mit y=1 eindeutig abtrennt.

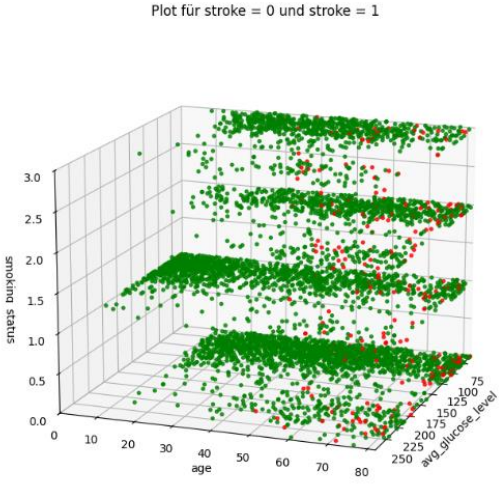
4.5 Visualisierung der Zerstreuung der Strokes im Raum

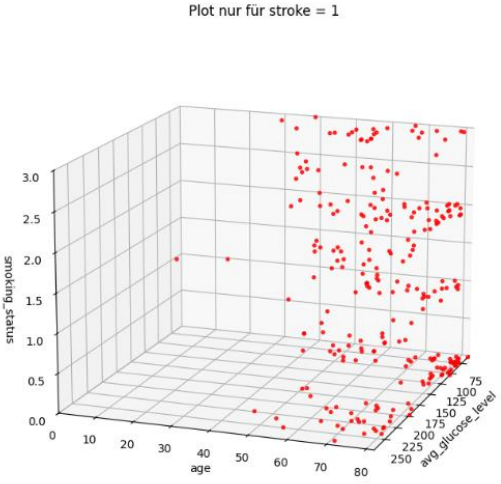
Wir haben festgestellt, dass wir wenig stroke-Daten haben, und diese im Datenraum wild verteilt sind. Das wollen wir hier nochmals visualisieren.

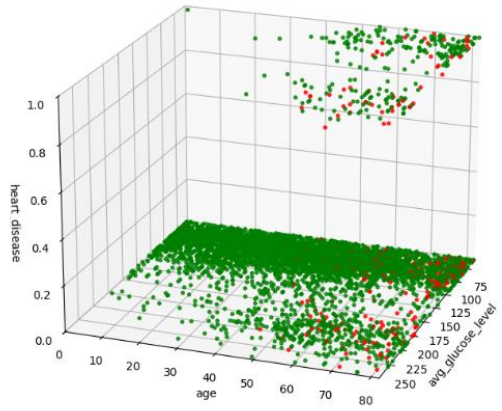
Wir plotten dafür jeweils zwei numerische Features (Alter und Blutzucker) auf den „Bodenachsen“ x1 und x2, und ein binäres Feature aus den Risikofaktoren (also Bluthochdruck, Herzkrankheit, Raucher) auf der vertikalen Achse x3. Rote Punkte sind dabei strokes, grüne Punkte sind nicht-strokes. Im Linken Plot ist jeweils der vollständige Datensatz zu sehen, recht jeweils nur die strokes.

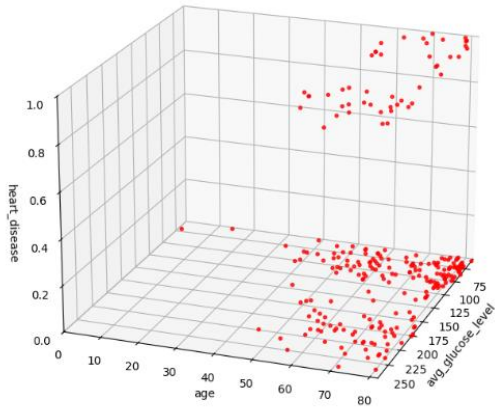












Die Plots machen nochmal deutlich:

- Keine der Risikogruppen hebt nur strokes nach oben ab
- Also liegen immer starke Mischungen vor. Das macht es sehr schwer, einzelne Strokes abzutrennen. Außerdem können wir auch keine Cluster gut abtrennen.

Allerdings kann man nochmal erkennen, dass sich die strokes aufs hohe Alter hin verdichten. Daher werden wir später entscheiden, dass wir nur ältere Personen durchsuchen, und junge Personen außen vor lassen.

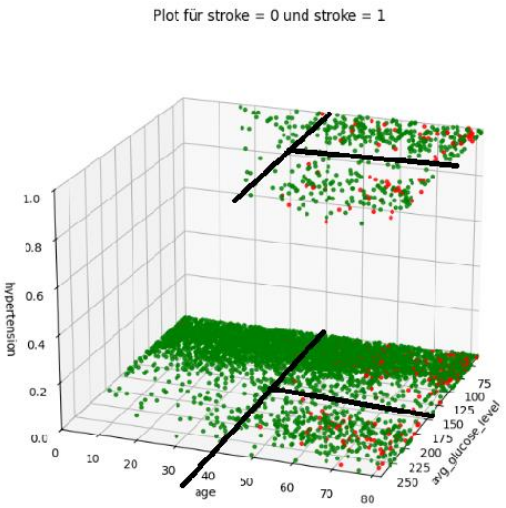
5. Fazit Datenstrukturen

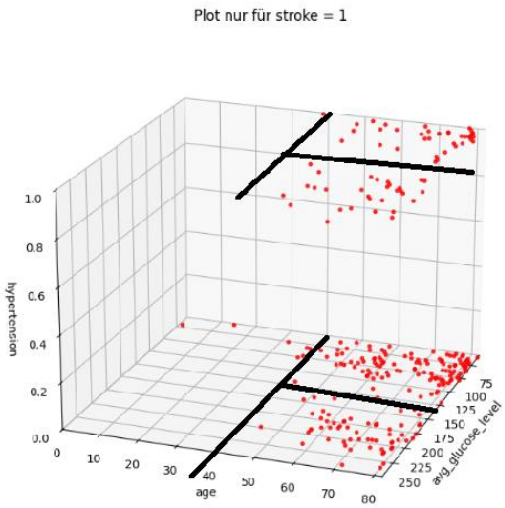
Wir fassen ein letztes Mal die Problemstellungen zusammen, vor die die Daten uns stellen:

- **Unbalancierte Klassen:** Strokes (y=1) sind in den Daten deutlich in der Unterzahl
- **Zerstreute Daten/“zerfranste Grenzen“:** Strokes sind nicht in gut isolierten Bereichen auffindbar, sondern in allen (!) Attributen stark unter die nicht-strokes gemischt!
- **Geringe Stroke-Dichte im Datenraum:** Dies ist eine Implikation aus den beiden Punkten oben. Wir werden uns das in Kapitel II 2.1 aber nochmal auf den Trainingsdaten vergegenwärtigen.
- **Risikogruppen:** Wir haben zwar Risikofaktoren identifizieren können (hohes Alter, Bluthochdruck, Herzkrankheit, Rauchen), aber keine davon war in den Daten so eklatant in den Strokes präsent, dass wir über den Risikofaktor große stroke-Cluster abtrennen könnten.

Dennoch haben wir festgestellt, dass im jungen Alter (ca. im Alter unter 50-55 Jahren) kaum strokes auftreten. Daher würde sich anbieten, diese Daten abzutrennen und nicht zu analysieren. Wir könnten uns also auf ältere Menschen spezialisieren.

Des Weiteren könnten wir die Gruppe mit hohem Blutzucker (Diabetiker / potenzielle Diabetiker) abtrennen (Clustering). Man erkennt im Plot unten nochmal, dass sich dort (links unten) zwei relativ gut abgrenzbare Cluster ergeben würden. Links unten ist die Übermacht grüner Punkte im Bereich avg_glucose_level <150 deutlich größer, bei > 150 etwas kleiner (also sind hier im Verhältnis mehr strokes, was wir schon gesehen haben). Eventuell ließen sich dann auf den Clustern verschiedene Algorithmen spezialisiert trainieren und verbessern lassen.





II. Datenvorbereitung

1. Versuchsweises Erzeugen von Hilfsfeatures

Um die schwierige Datenlage versuchsweise zu verbessern, versuchen wir, einige neue Spalten zu erzeugen (basierend auf den Originalspalten), die eventuell dabei helfen, Datenbereiche, die nur (oder zumindest anteilig relativ viele) strokes enthalten, abzutrennen.

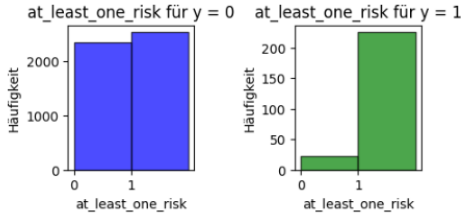
Dabei ist die Idee, dass wir die Risikofaktoren (Bluthochdruck, Herzkrankheit, Rauchen, Alter) kombinieren. Eventuell finden wir so „Hochrisikogruppen“, bei denen sich die strokes stark häufen (y=1), und bei denen kaum nicht-strokes enthalten (y=0) sind.

Wir erzeugen also folgende Spalten:

- **Age_above_60**: (haben wir erzeugt, bevor wir entschieden haben, den cut bei 55 zu machen; die Spalte ist also mittlerweile obsolet)
- **High_glucose**: Blutzuckerwert > 150 (also potenzieller Diabetiker), dann 1; sonst 0
- **Did_smoke**: Zusammenfassung ehemaliger und aktiver Raucher (beide bekommen hier 1, alle anderen 0)
- **Heart_risk**: 1, wenn Bluthochdruck oder Herzkrankheit, sonst 0
- **At_least_one_risk**: Mindestens eines aus [Bluthochdruck, Herzkrankheit] erfüllt, dann 1, sonst 0
- **At_least_one_risk_and_high_age**: At least one risk, und zusätzlich älter 40 (wurde erzeugt, mittlerweile ebenfalls obsolet)
- **All_risks**: 1, wenn ALLE Risikofaktoren (Herzkrank, Bluthochdruck, Raucher, hohes Alter) erfüllt sind, 0 sonst
- **Risk_sum**: Summe aller Risiken (Herzkrank, Bluthochdruck, Raucher, hohes Alter)

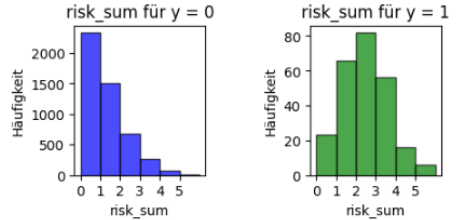
„Leider“ (im Sinne unserer Machine Learning Fragestellung) zeigt sich auch hier, dass keines der neuen Features eine hohe bedingte Wahrscheinlichkeit auf einen stroke hat (auch wenn das Vorhandensein einer 1 des Features die Wahrscheinlichkeit erhöht). Es findet also auch hier nirgends gute Abgrenzung statt.

Wir zeigen hier nur die Plots für „at_least_one_risk“ und „risk_sum“:



Crosstable

stroke	0	1	category_sum	share_category_of_total	healthy-%	stroke-%
at_least_one_risk						
0	2333	23	2356	46.1%	99.0%	1.0%
1	2528	226	2754	53.9%	91.8%	8.2%

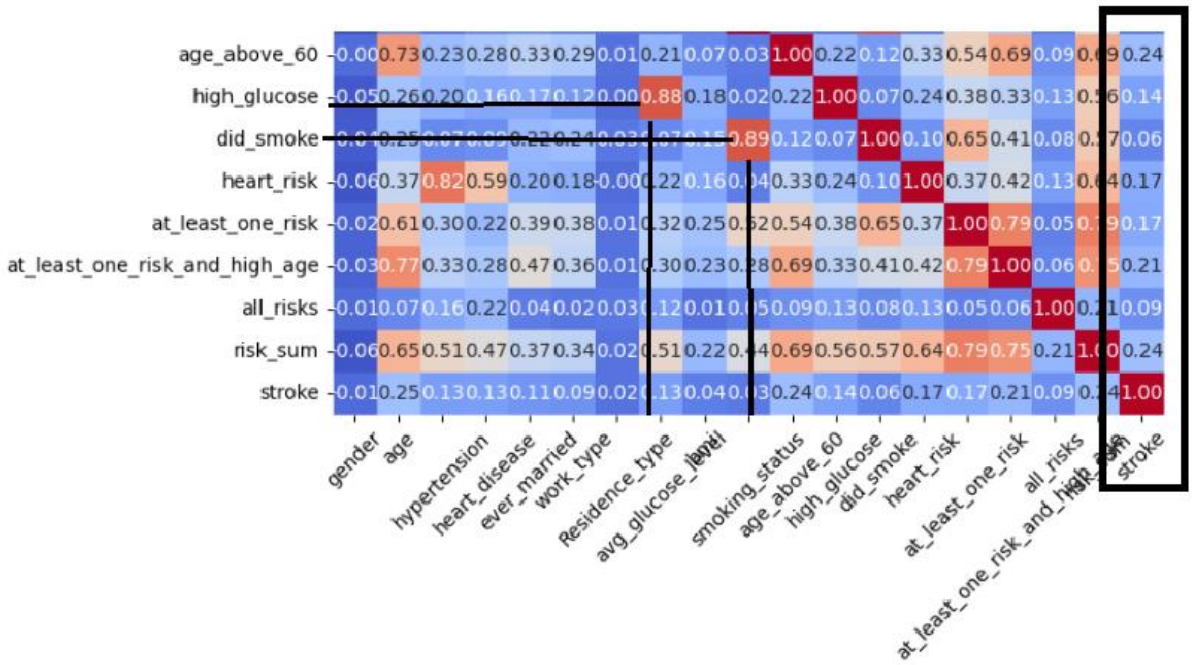


Crosstable

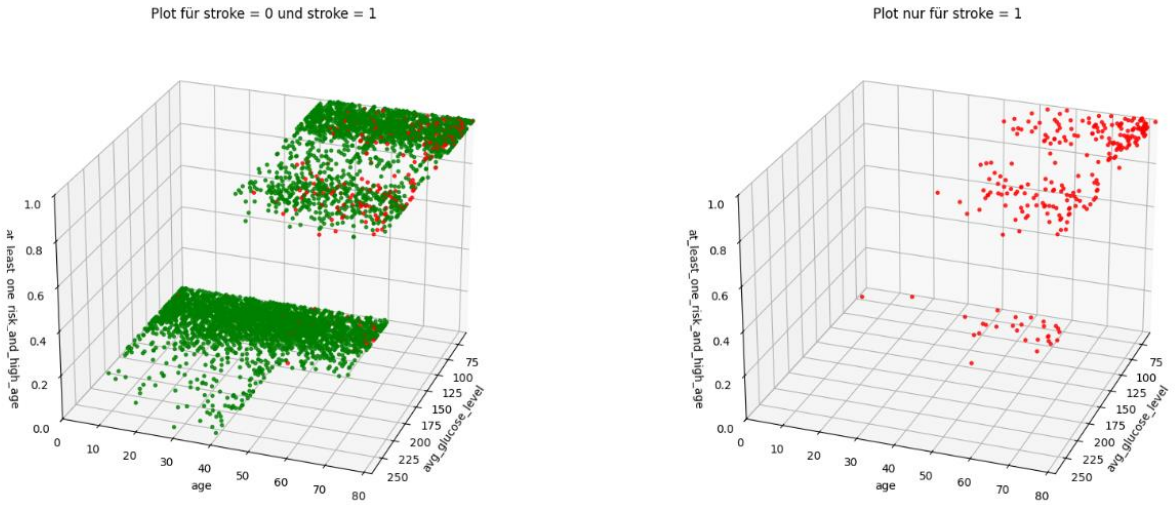
stroke	0	1	category_sum	share_category_of_total	healthy-%	stroke-%
risk_sum						
0	2333	23	2356	46.1%	99.0%	1.0%
1	1498	66	1564	30.6%	95.8%	4.2%
2	684	82	766	15.0%	89.3%	10.7%
3	267	56	323	6.3%	82.7%	17.3%
4	71	16	87	1.7%	81.6%	18.4%
5	8	6	14	0.3%	57.1%	42.9%

Aus medizinischer Sicht sind die Kreuztabellen sehr interessant („risk_sum“ = 5 hat 42,9% bedingte Wahrscheinlichkeit für einen stroke!!), ABER leider betrifft das in unseren Daten nur sehr wenige Datenpunkte (bei risk_sum = 5 nur 6 von 14 Punkten...).

Leider erzeugt auch keines der neuen Features eine hohe Korrelation mit den strokes, die man im Ausschnitt der erweiterten Korrelationsheatmap erkennen kann. Zudem dürften die Features nicht im Kombination mit den Originalfeatures verwendet werden, da sie zu diesen teilweise stark korrelieren (siehe did_smoke und smoking_status, oder auch risk_sum und alle Risikofaktoren). Als interessante Nebenbeobachtung ist high_glucose (>150) stark korreliert mit dem Residence_type (0= rural, 1= urban).



Wir betrachten exemplarisch noch die Kombispalte aus at_least_one_risk und hohem Alter:



Dieser Plot hat uns letzten Endes dazu bewegt, uns aufs Alter Ü55 zu konzentrieren.

Wir stellen am Ende der Datenvorbereitung diese Zusatzfeatures trotzdem im Datensatz zur Verfügung (auch wenn sie doch nicht so vielversprechend sind, wie wir gehofft hatten), für den Fall, dass sie bei einzelnen Algorithmen ausprobiert werden sollten.

2. Ausreißer, Dateneingrenzung, Resampling

Im Verlauf der Datenanalyse oben haben wir festgestellt:

- Es gibt zwei Ausreißer (Alter 1 Jahr und 14 Jahre) die wir entfernen wollen.
- Es gibt eine Person mit Gender „Other“ (ohne stroke), diese wollen wir auch entfernen.

Da wir aber nach weiterer Überlegung festgestellt haben, dass wir nur Personen ab höherem Alter untersuchen wollen (jünger gibt es viel zu wenig strokes in den Daten), beschließen wir ohnehin, alle Daten mit Alter < 55 Jahre zu verwerfen.

Außerdem haben wir die Spalten „ever_married“ und „work_type“ verworfen.

Da wir uns nun auf Personen mit Alter >= 55 beschränkt haben, hoffen wir auf dichtere Nachbarschaften (da im hohen Alter mehr strokes auftreten).

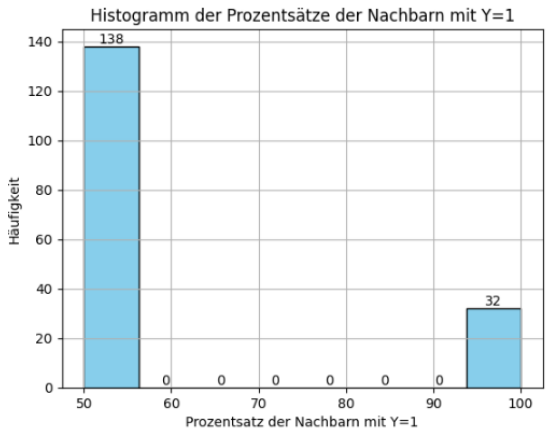
Wir schauen uns kurz die resultierende stroke-Dichte in Nachbarschaften in X_train an.

2.1 Kurze Analyse der Nachbarschaften von Strokes

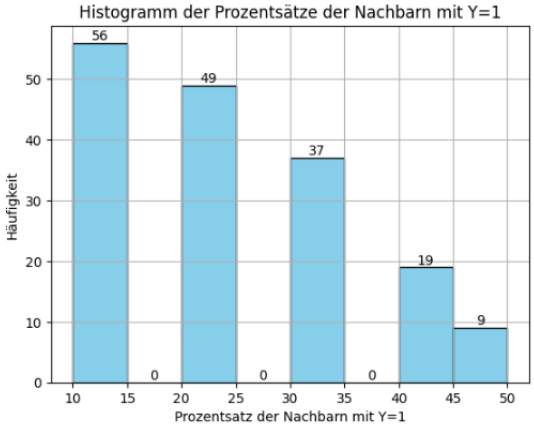
Wir betrachten nun noch die Nachbarschaften der strokes im Trainingsdatenraum:

- Für Punkte mit y=1 berechnen wir die N nächsten Nachbarn
- In diesen Nachbarschaften berechnen wir den Prozentsatz anderer vorhandener Strokes
- In den folgenden Histogramm sehen wir für N=2 und N=10 dann die Häufigkeitsbalken, wie oft strokes N-Nachbarschaften mit X% (x-Achse) strokes haben.

2 nächste Nachbarn



10 nächste Nachbarn



Wir sehen also:

- Wenn es nur um die 2 nächsten Nachbarn von strokes geht, haben gerade mal 47 von 249 Strokes selbst wiederum 100% strokes in ihrer Nachbarschaft. 167 von 249 strokes haben keinen stroke unter den nächsten 2 Nachbarn. Das heißt: **Die meisten strokes sind trotzdem noch von nicht-strokes umgeben**!
- Wenn es um die nächsten 10 Nachbarn von strokes geht, haben wir nur noch Nachbarschaften, die maximal 50% strokes enthalten, meist noch deutlich weniger.

Unter den nächsten 10 Nachbarn wird noch deutlicher, dass wir quasi nirgends eine hohe Dichte an strokes haben.2.2 Resampling für Optimierungsversuche

Wir versuchen außerdem

3. Split der Daten in Trainings- und Testdaten

Wir nehmen eine Abtrennung von 20% der Daten als Testdaten mithilfe von train_test_split vor, wobei wir nach den strokes (also der Zielspalte) stratifizieren.

4. Auffüllen fehlender Werte

Beim BMI haben wir gesehen, dass ca. 200 Werte fehlen. Wir füllen diese NACH dem Split in Trainingsdaten und Testdaten mit dem Durchschnitt ders Trainingsdaten-BMI auf.

Die Entscheidung für den Durchschnitt fällen wir, weil der BMI

- Keine hohe Korrelation zu strokes hat, und wir hier keine Energie in anspruchsvolleres Imputing stecken wollten
- Der BMI einigermaßen normalverteilt (mit leichter rechtsschiefe) ist.

5. Skalieren

Wir haben nicht skaliert. Wir werden Algorithmen verwenden, die teilweise evt. untershciedliche Skalierungen bevorzugen, daher überlassen wir das Skalieren dem Arbeittschritt der Einzelanalyse/-vorbereitung des jeweiligen Algorithmus

III. Zusammenfassung der Ideen zur Datenverbesserung

Um die vielfach beschriebenen Schwierigkeiten der Datenlage anzugehen, sehen wir folgende Möglichkeiten:

- **Feature-Engineerung** zur Abtrennung von stroke-reichen Clustern:
  - Siehe beschriebene Versuche, kaum erfolgreich. Ursache für das Scheitern: Strokes liegen immer von nicht-strokes umgeben. Quasi unmöglich, in größerem Stil strokes abzusondern.
- **Datenteile weglassen** (hier haben wir eine Einschränkung auf das Alter Ü55 vorgenommen, und einen Teil an irrelevanten Daten (jüngere Menschen, wenig Risikofaktoren, keine strokes) weggelassen

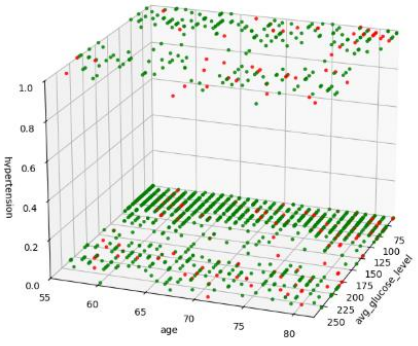
Eine weitere Möglichkeit wäre **Resampling:**

- Oversampling (SMOTE), um den Anteil an Strokes auf 50% „aufzupumpen“
  - Problem: Wie genau wird gesampelt?
- Undersampling: Wegstreichen von nicht-strokes, bis ein 50-50-Verhältnis erreicht ist
  - Problem: Danach zu wenig Daten übrig

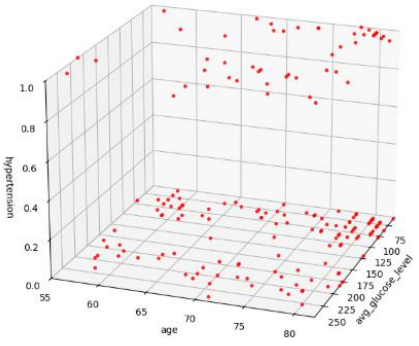
Nach intensiver Diskussion haben wir uns entschieden, SMOTE nicht zu verwenden, aus folgendem Grund:

- Smote sucht sich für einen stroke die n nächsten strokes (n ist Parameter von smote), und erzeugt Linearkombinationen zwischen diesen n Nachbarn.
- Das heißt:
  - Smote eignet **sich gut, um bestehende Inseln zu verdichten**!
  - Smote ist aber **schlecht für weit verteilte, und dünn ausgesäte Einzelelemente** der Minderheitsklasse! Smote erzeugt dann auf den Strecken

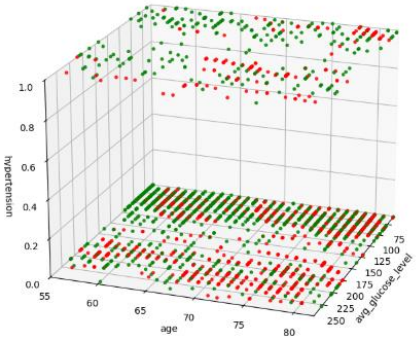
zwischen bestehenden strokes neue strokes, die häufig in Bereiche fallen, in denen sonst nur nicht-strokes lägen! Zur Visualisierung sehen wir unten nochmal einen Plot ohne Smote (obige zwei Plots) und nach Smote (untenstehende zwei Plots).

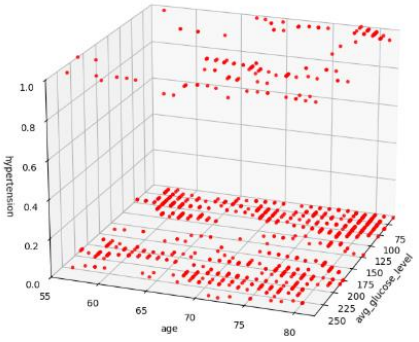


Plot für stroke = 0 und stroke = 1



Plot nur für stroke = 1

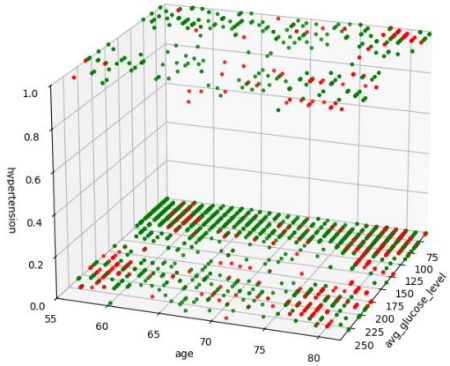




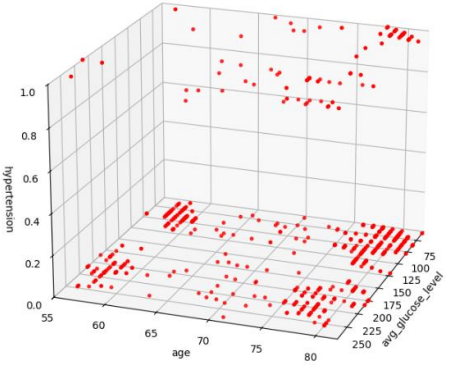
Smote bildet synthetische Daten in der konvexen Hülle der Minderheitsklasse, und zwar ohne Rücksicht auf verschiedene Dichten dort. Um das Resampling in Minderheitsteilmengen mit besonders geringer Dichte zu vermeiden, könnten wir SMOTE z.B. nur auf explizit ausgewählte Teilmengen anwenden, zum Beispiel zur eine bestimmte Altersklasse (oder auch eine Teilmenge der Risikopatienten „at_least_one_risk“:

```
X_train_filtered = X_train[((X_train["age"]>=55)&(X_train["age"]<=60)) | (X_train["age"]>=75)]
Y_train_filtered = Y_train.loc[X_train_filtered.index]
```

Plot für stroke = 0 und stroke = 1



Plot nur für stroke = 1



Damit erhalten wir also keine großflächige Verdichtung, sondern nur eine in dedizierten Teilbereichen.

Weiters: Eine PCA oder ein Clustering wurden nicht durchgeführt, wäre aber noch eine Option zur vertieften Analyse.

Es könnte noch die Spalte „Residence_type“ verworfen werden. Dies haben wir aber erst festgestellt, als bereits Einzelanalysen der Algorithmen gestartet waren, und wir den Datensatz nicht nochmals anpassen wollten.