INCOME CLASSIFICATION

Drakaki, Nikolitsa
Freter, Freter
Symhoven, Simon

28. Juni 2023

AGENDA

- Einleitung
- Datensatz
  - Deskriptive Statistik
  - Preprocessing
- PCA
- Clusteranalyse
- Logit Modell
- Random Forest
- Neuronales Netz
- Modell Vergleich

EINLEITUNG

Die Bedeutung sozioökonomischer Faktoren

Wir untersuchen, wie verschiedene Faktoren wie Bildung, Beruf, Alter und mehr das Einkommen einer Person beeinflussen.

DATENSATZ

Income Classification

INCOME CLASSIFICATION

- Vorhersage des Einkommens
  - < 50k $
  - > 50k $
- Insgesamt 48.842 Personen
- 14 Variablen inkl. Einkommen
- 6 numerische Variablen
- 8 kategoriale Variablen inkl. Einkommen

```
workclass contains 8 labels
['State-gov' 'Self-emp-not-inc' 'Private' 'Federal-gov' 'Local-gov'
 'Self-emp-inc' 'Without-pay' 'Never-worked']
education contains 16 labels
['Bachelors' 'HS-grad' '11th' 'Masters' '9th' 'Some-college' 'Assoc-acdm'
 'Assoc-voc' '7th-8th' 'Doctorate' 'Prof-school' '5th-6th' '10th'
 '1st-4th' 'Preschool' '12th']
marital-status contains 7 labels
['Never-married' 'Married-civ-spouse' 'Divorced' 'Married-spouse-absent'
 'Separated' 'Married-AF-spouse' 'Widowed']
occupation contains 14 labels
['Adm-clerical' 'Exec-managerial' 'Handlers-cleaners' 'Prof-specialty'
 'Other-service' 'Sales' 'Craft-repair' 'Transport-moving'
 'Farming-fishing' 'Machine-op-inspct' 'Tech-support' 'Protective-serv'
 'Armed-Forces' 'Priv-house-serv']
relationship contains 6 labels
['Not-in-family' 'Husband' 'Wife' 'Own-child' 'Unmarried' 'Other-relative']
race contains 5 labels
['White' 'Black' 'Asian-Pac-Islander' 'Amer-Indian-Eskimo' 'Other']
sex contains 2 labels
['Male' 'Female']
native-country contains 41 labels
['United-States' 'Cuba' 'Jamaica' 'India' 'Mexico' 'South' 'Puerto-Rico'
 'Honduras' 'England' 'Canada' 'Germany' 'Iran' 'Philippines' 'Italy'
 'Poland' 'Columbia' 'Cambodia' 'Thailand' 'Ecuador' 'Laos' 'Taiwan'
 'Haiti' 'Portugal' 'Dominican-Republic' 'El-Salvador' 'France'
 'Guatemala' 'China' 'Japan' 'Yugoslavia' 'Peru'
 'Outlying-US(Guam-USVI-etc)' 'Scotland' 'Trinadad&Tobago' 'Greece'
 'Nicaragua' 'Vietnam' 'Hong' 'Ireland' 'Hungary' 'Holand-Netherlands']
```

INCOME CLASSIFICATION

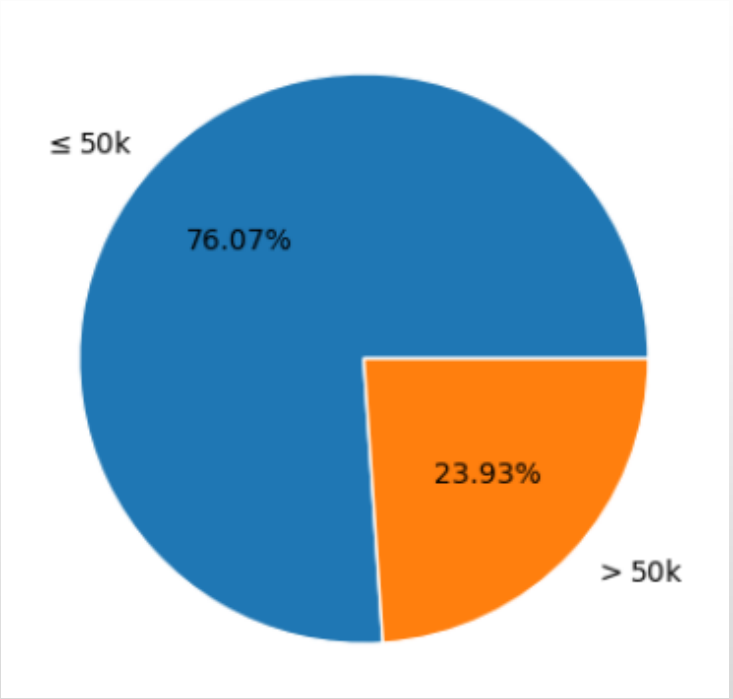
Datensatz

```
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             48842 non-null  int64
 1   workclass       46043 non-null  object
 2   fnlwgt          48842 non-null  int64
 3   education       48842 non-null  object
 4   marital-status  48842 non-null  object
 5   occupation      46033 non-null  object
 6   relationship    48842 non-null  object
 7   race            48842 non-null  object
 8   sex             48842 non-null  object
 9   capital-gain    48842 non-null  int64
 10  capital-loss    48842 non-null  int64
 11  hours-per-week  48842 non-null  int64
 12  native-country  47985 non-null  object
 13  income          48842 non-null  int64
dtypes: int64(6), object(8)
memory usage: 5.2+ MB
```

NaN's wurden durch den mode ersetzt

age	0	age	0
workclass	2799	workclass	0
fnlwgt	0	fnlwgt	0
education	0	education	0
marital-status	0	marital-status	0
occupation	2809	occupation	0
relationship	0	relationship	0
race	0	race	0
sex	0	sex	0
capital-gain	0	capital-gain	0
capital-loss	0	capital-loss	0
hours-per-week	0	hours-per-week	0
native-country	857	native-country	0
income	0	income	0
dtype: int64		dtype: int64	

Verteilung des Einkommens



INCOME CLASSIFICATION

Datensatz

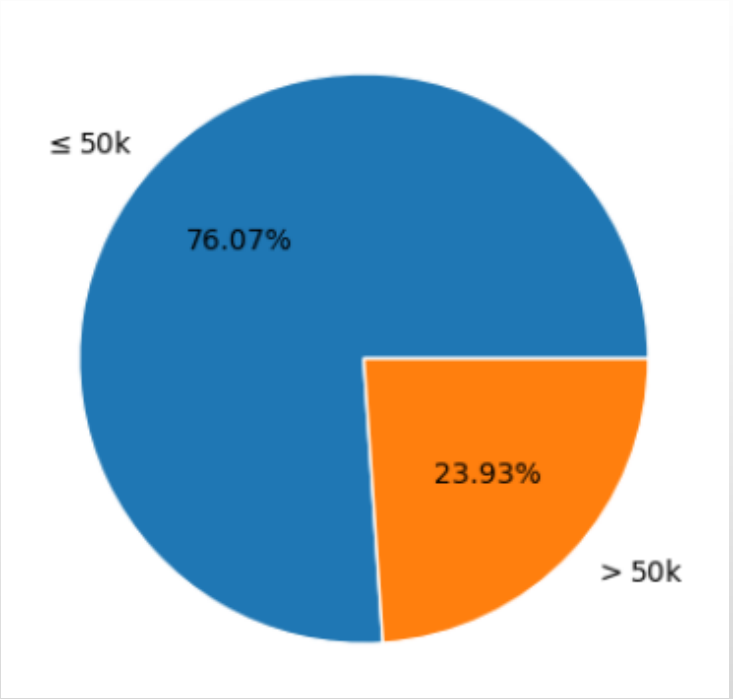
```
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             48842 non-null  int64
 1   workclass       46043 non-null  object
 2   fnlwgt          48842 non-null  int64
 3   education       48842 non-null  object
 4   marital-status  48842 non-null  object
 5   occupation      46033 non-null  object
 6   relationship    48842 non-null  object
 7   race            48842 non-null  object
 8   sex             48842 non-null  object
 9   capital-gain    48842 non-null  int64
 10  capital-loss    48842 non-null  int64
 11  hours-per-week  48842 non-null  int64
 12  native-country  47985 non-null  object
 13  income          48842 non-null  int64
dtypes: int64(6), object(8)
memory usage: 5.2+ MB
```

NaN's wurden durch den mode ersetzt

```
age                0
workclass       2799
fnlwgt             0
education          0
marital-status     0
occupation      2809
relationship       0
race               0
sex                0
capital-gain       0
capital-loss       0
hours-per-week     0
native-country   857
income             0
dtype: int64
```

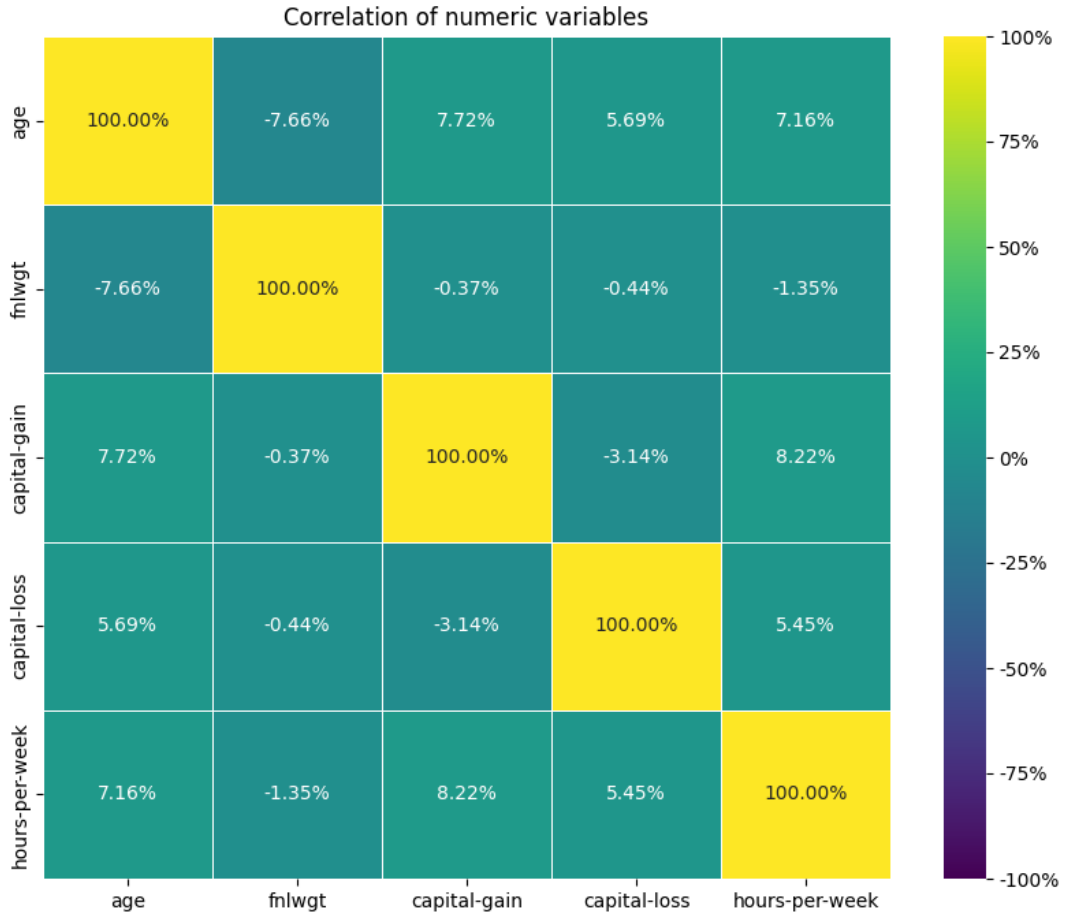
```
age                0
workclass          0
fnlwgt             0
education          0
marital-status     0
occupation         0
relationship       0
race               0
sex                0
capital-gain       0
capital-loss       0
hours-per-week     0
native-country     0
income             0
dtype: int64
```

Verteilung des Einkommens



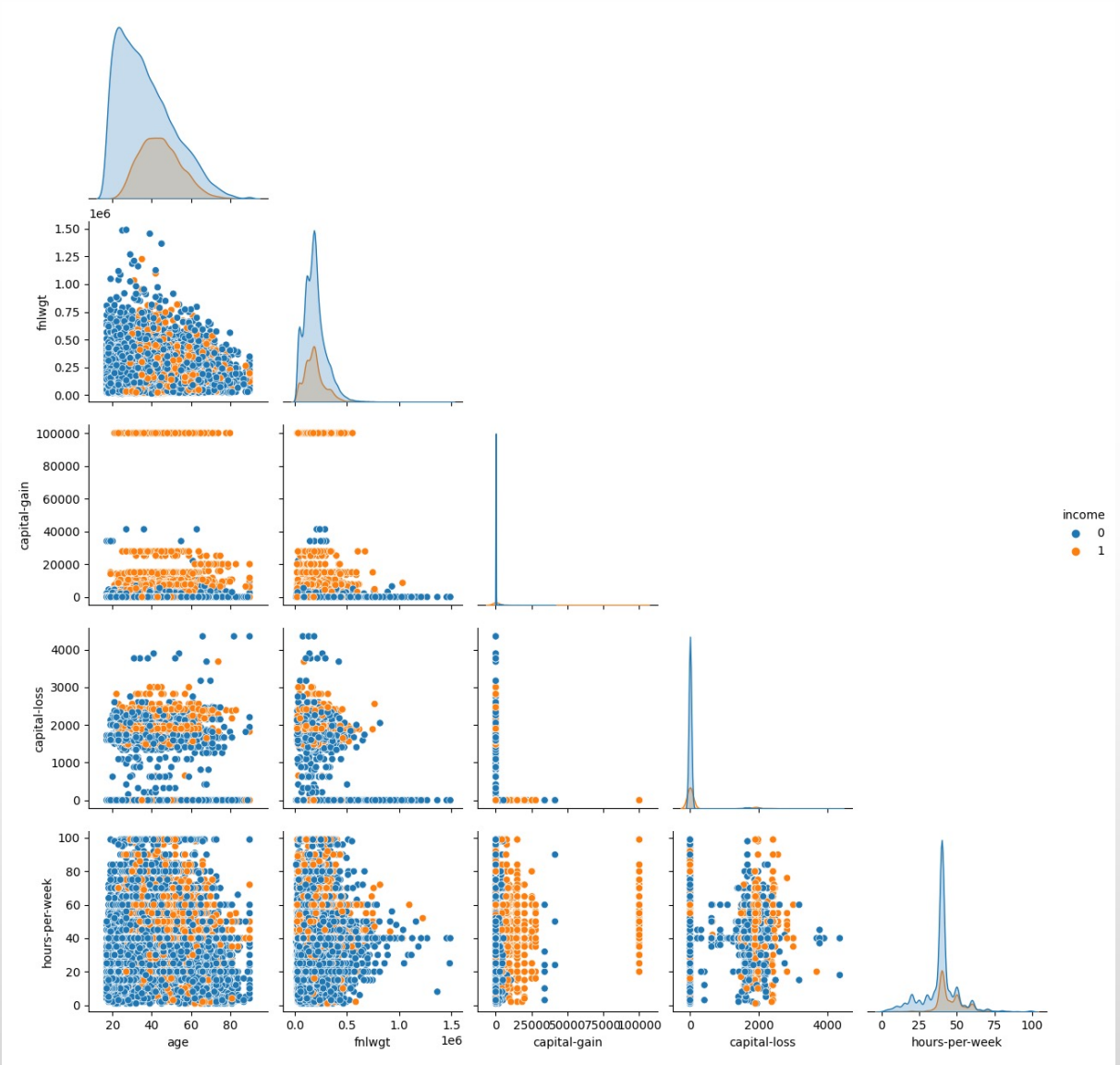
KORRELATION
DER NUMERISCHEN DATEN

- Keine Korrelation innerhalb des Datensatzes



VERTEILUNGEN
DER NUMERISCHEN DATEN

- Orange: > 50k €
- Blau: < 50k €

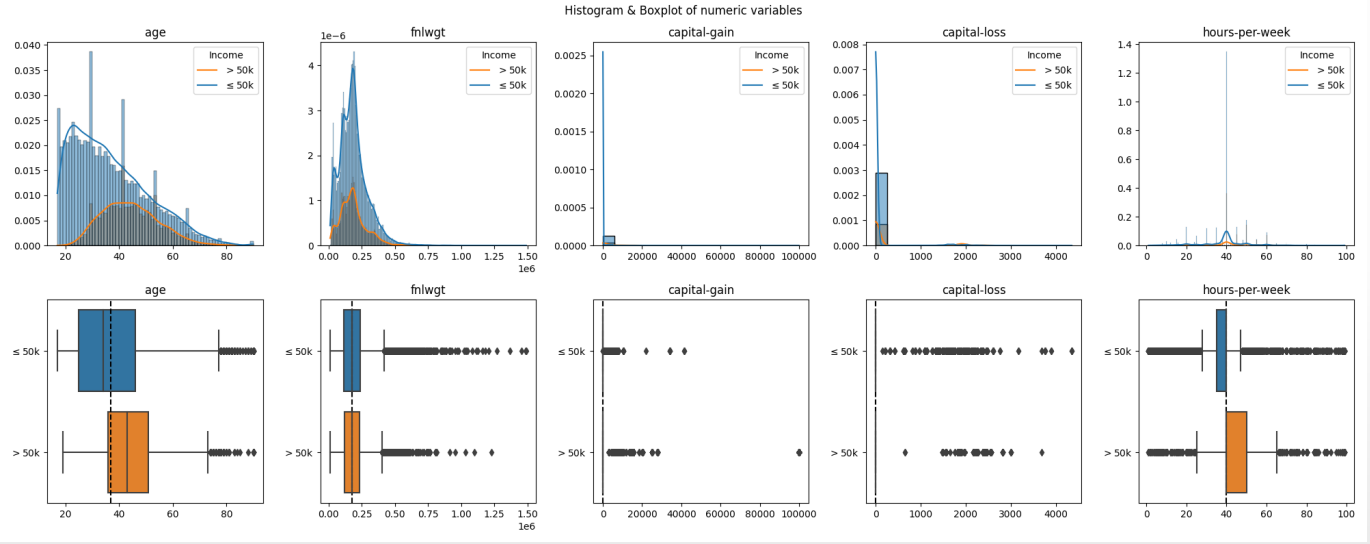


VERTEILUNGEN
DER NUMERISCHEN DATEN

- Orange: > 50k €
- Blau: < 50k €

<u>Idee:</u>

Logarithmische Transformation auf age und fnlwgt



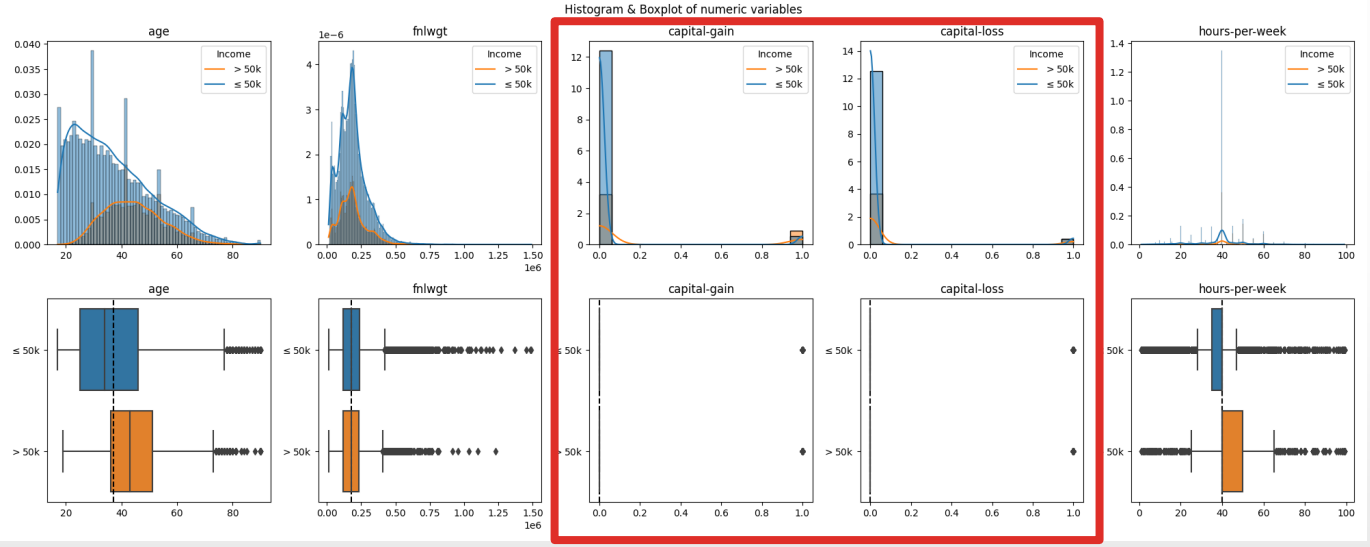
VERTEILUNGEN
DER NUMERISCHEN DATEN

<u>Ansätze:</u>

- Capping der Ausreißer (über 1000)

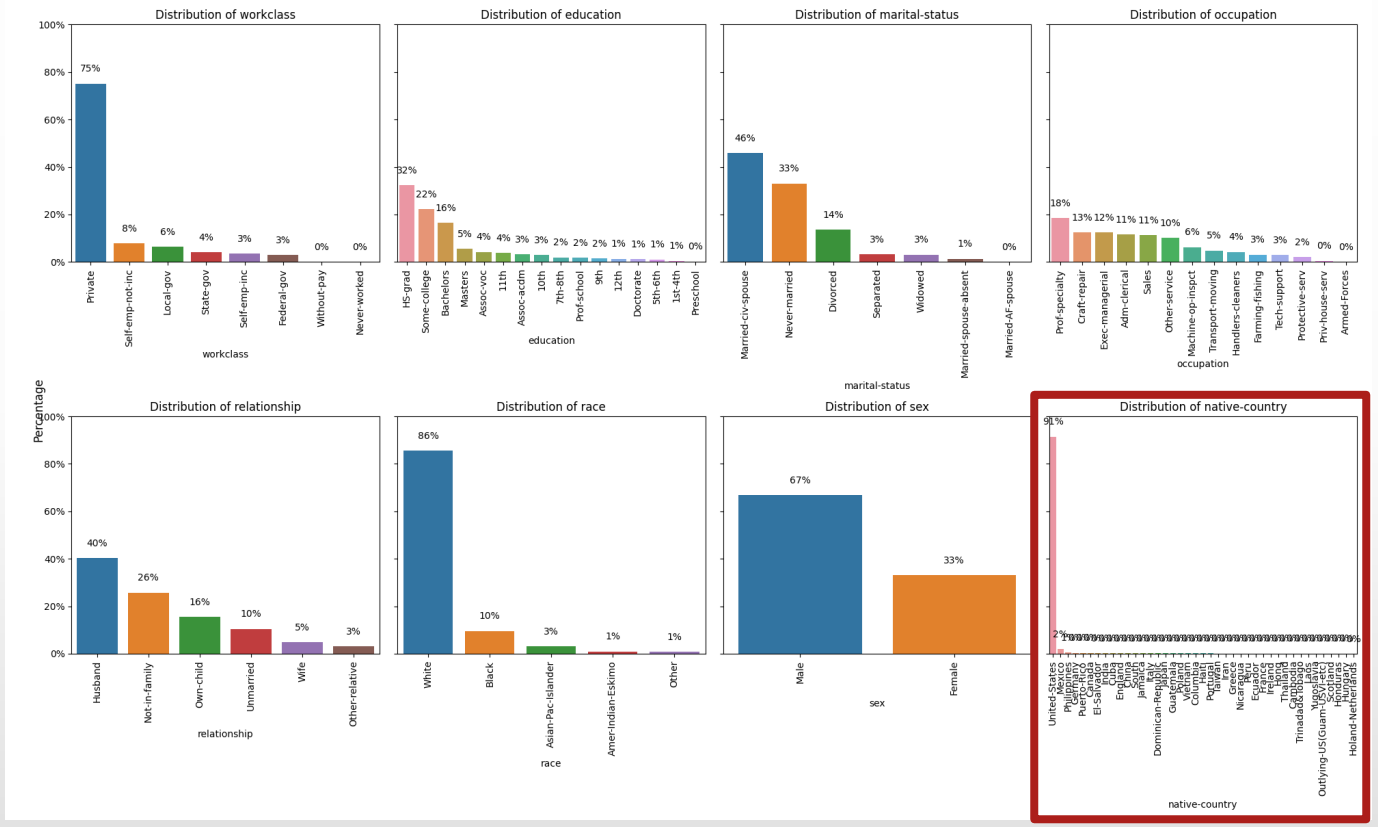
<u>Lösung:</u>

- Binäre Codierung
- Logarithmische Transformation auf age und fnlwgt (für nicht lineare Transformation)



VERTEILUNGEN
DER KATEGORIALEN DATEN

- Absolute Verteilung der Features
- Überwiegend US Staatsbürger
- Native country lassen wir fallen

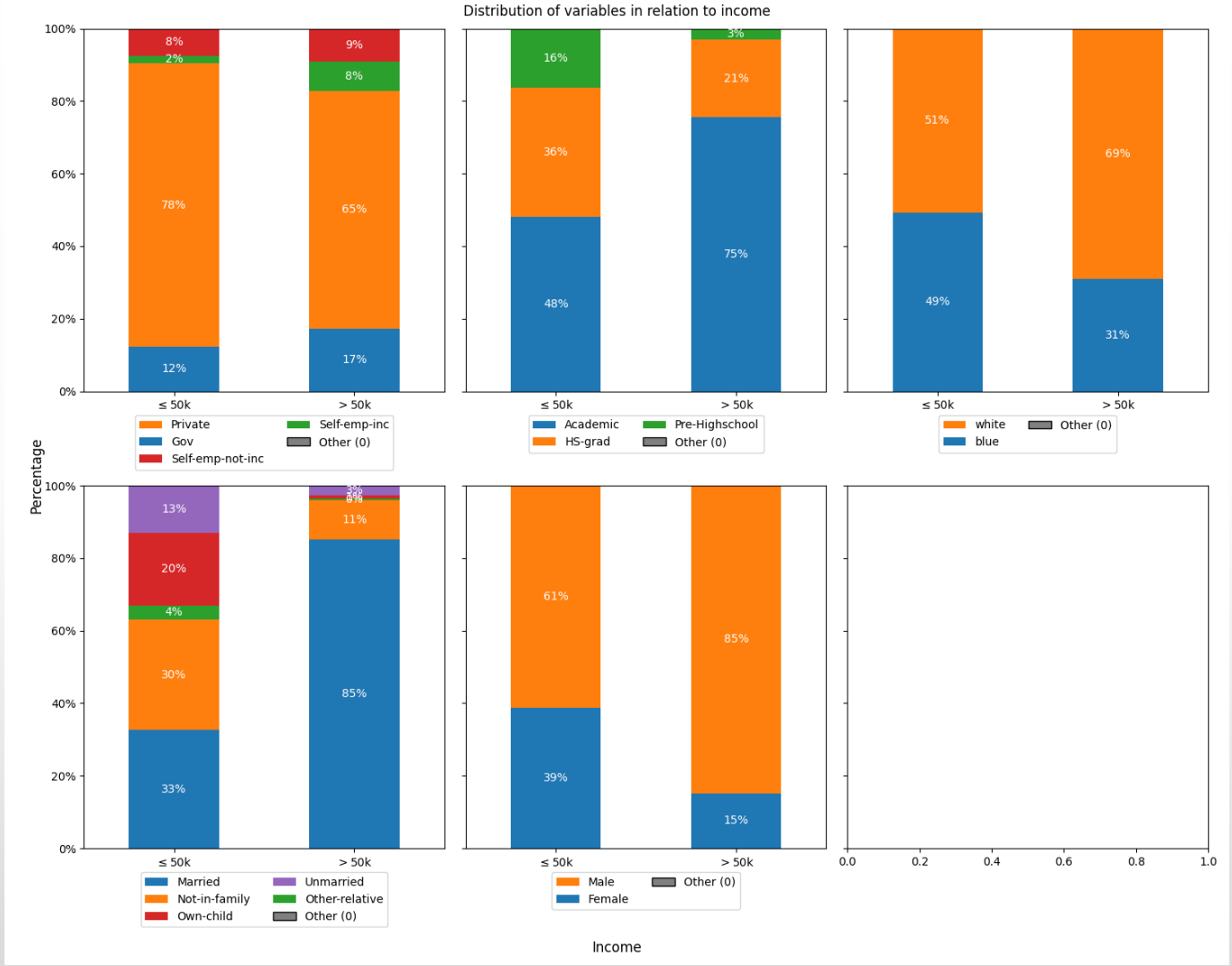


VERTEILUNGEN
DER KATEGORIALEN DATEN

- Relative Verteilung in Bezug auf das Einkommen
- Native country besteht zu über 90% aus einem Land
  - sowohl für Einkommen über als auch unter 50k $
- kategorialen Daten zusammenzufassen
  - Stichwort: Separierbarkeit

VERTEILUNGEN
DER KATEGORIALEN DATEN

- Features wurden zusammengefasst
- Einige Features wurden entfernt
  - Native country
  - Race
  - Marital status
- OneHot Kodierung der kategorialen Features
- Min-Max Skalierung aller Features



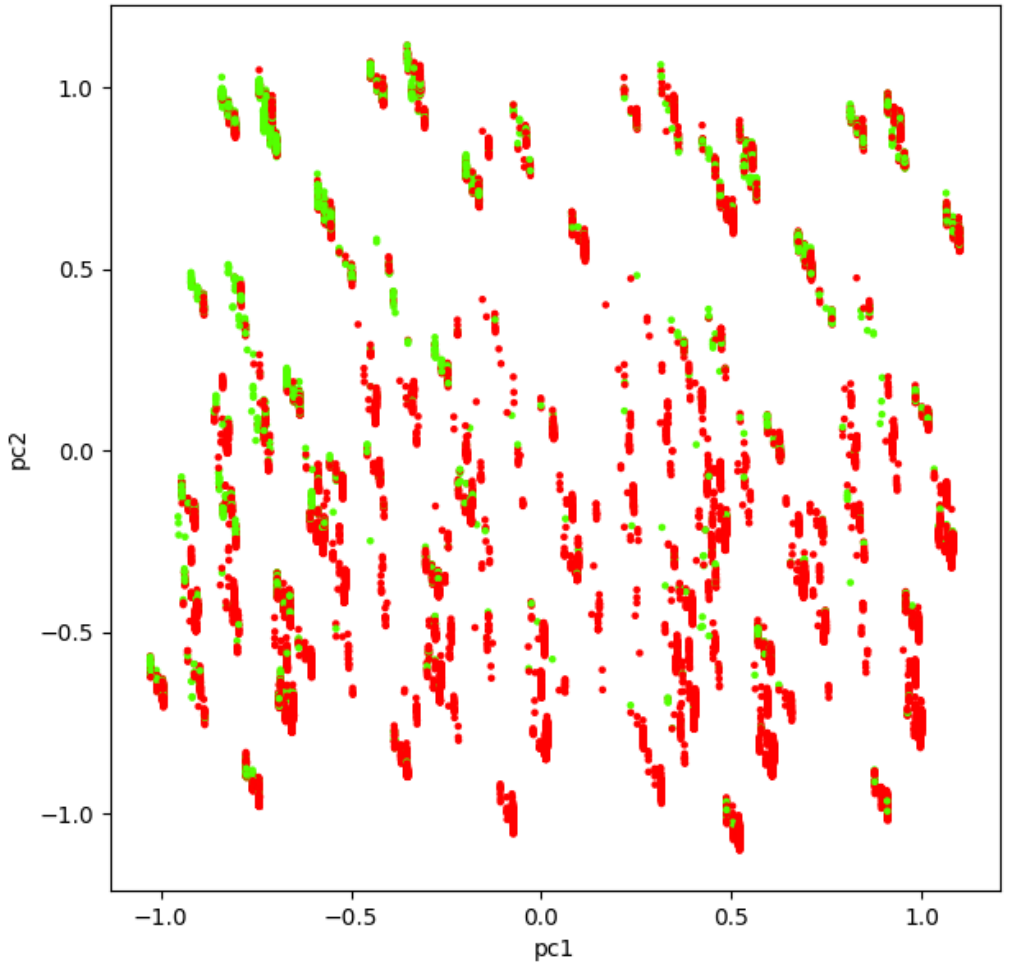
FINALE FEATURES

```
workclass contains 4 labels
['Gov' 'Self-emp-not-inc' 'Private' 'Self-emp-inc']
education contains 3 labels
['Academic' 'HS-grad' 'Pre-Highschool']
occupation contains 2 labels
['blue' 'white']
relationship contains 5 labels
['Not-in-family' 'Married' 'Own-child' 'Unmarried' 'Other-relative']
sex contains 2 labels
['Male' 'Female']
```

HAUPTKOMPONENTEN ANALYSE

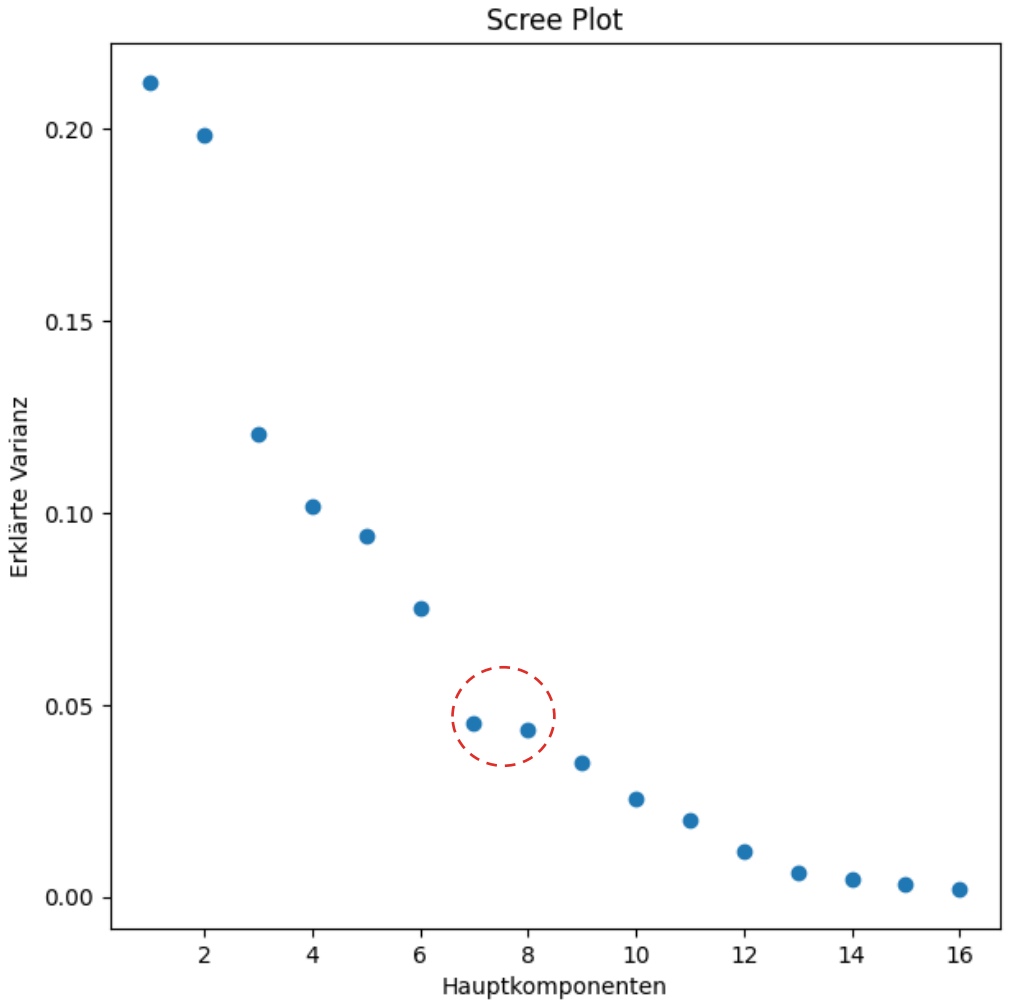
HAUPTKOMPONENTEN ANALYSE

- Dimensionsreduktion zur Vereinfachung des Datensatzes



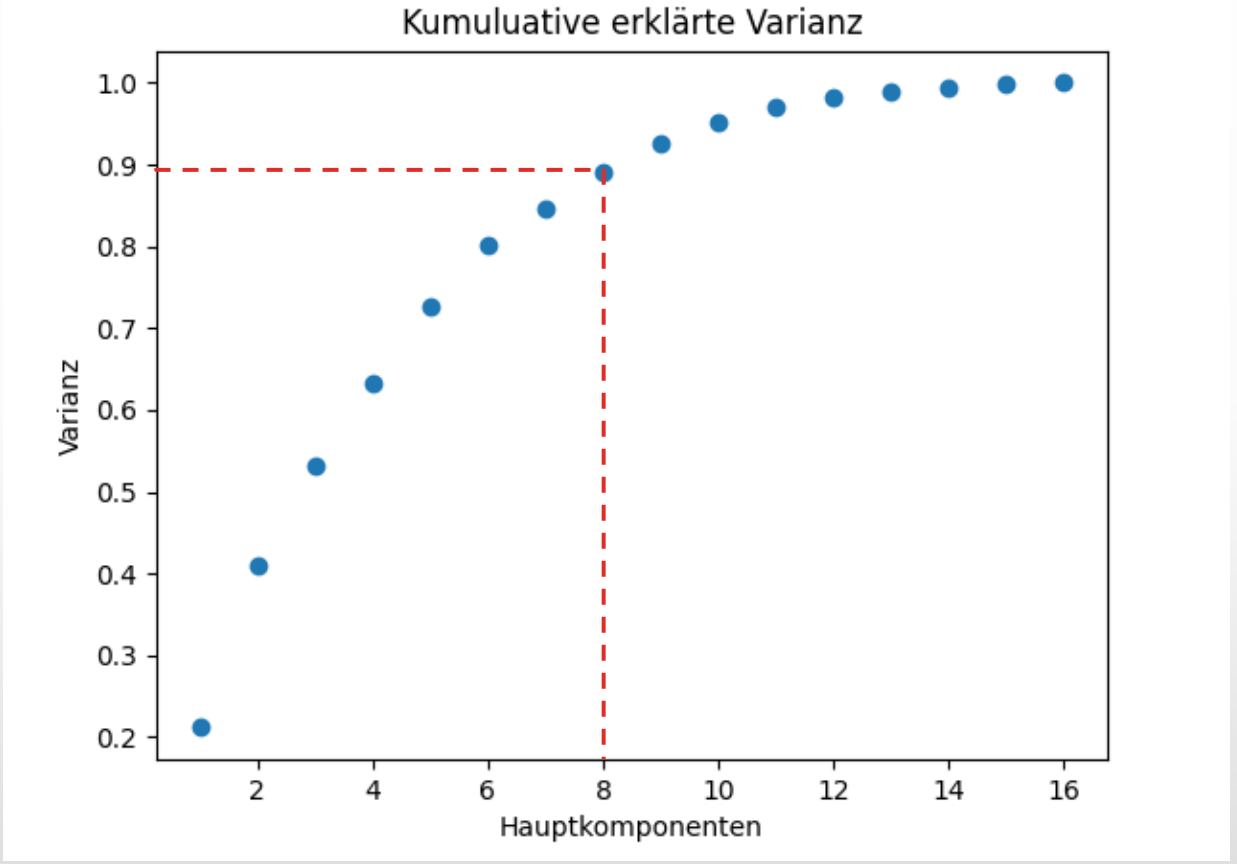
SCREE PLOT

- Erklärte Varianz pro Hauptkomponente
- Ellbogen-Methode
  - 7 oder 8 Hauptkomponenten scheinen sinnvoll



HAUPTKOMPONENTEN ANALYSE

- Maximierung der Varianz im Datensatz bei gleichzeitiger Dimensionsreduktion
- 8 Hauptkomponenten erklären fast 90% der Varianz



CLUSTERANALYSE

CLUSTERING

Clusteranalyse

- Unsupervised Machine-Learning-Technik
- Visualisierungsmethode

K-Means Clustering

- Ein einfaches Verfahren zur Partitionierung eines Datasets in K unterschiedliche, sich nicht überschneidende Cluster, deren Anzahl vordefiniert ist. Der Gedanke dahinter ist, dass ein gutes Clustering eine möglichst geringe Varianz innerhalb der Cluster aufweist.

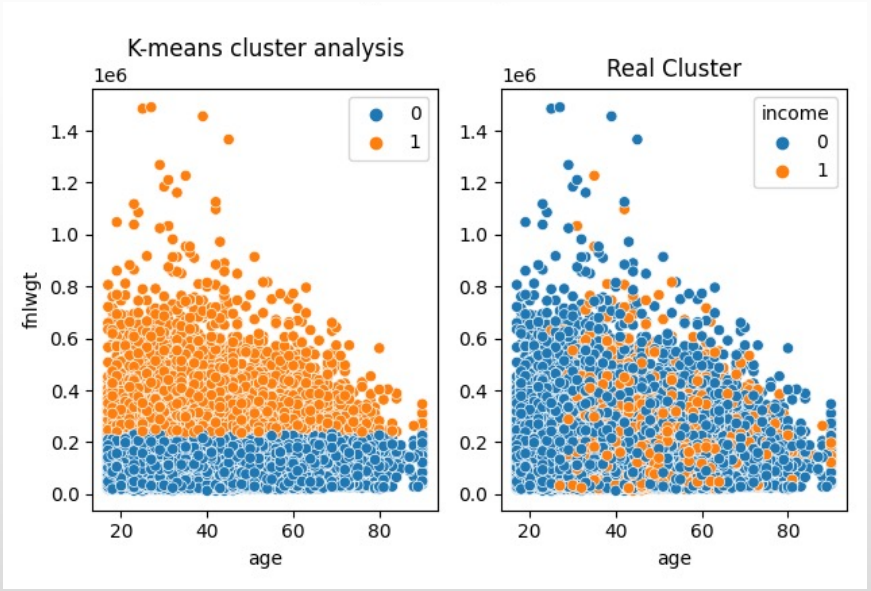
Warum haben wir die K-Mean Clustering Methode ausgewählt?

- Sehr beliebter und gut erforschter Algorithmus
- Verständliches Verfahren mit einer schnellen Implementierung
- Am häufigsten verwendbares Verfahren

K-MEANS ANALYSE

- K-Means kann nur bei den numerischen Variablen angewendet werden
- Scatterplots: Die Punkte sind gemäß dem Einkommen gefärbt

Age vs. financial weight



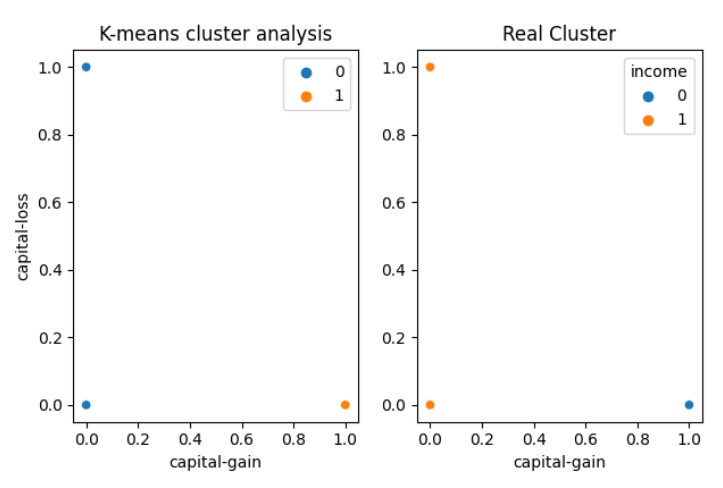
K-MEANS ANALYSE

- K-Means Plots: Die Punkte sind von der K-Means in zwei Gruppen aufgeteilt
- Die zwei Gruppen der K-Means Analyse entsprechen bei allen Paaren der nummerischen Variablen die echten Werte nicht perfekt. Es gibt keine Cluster.

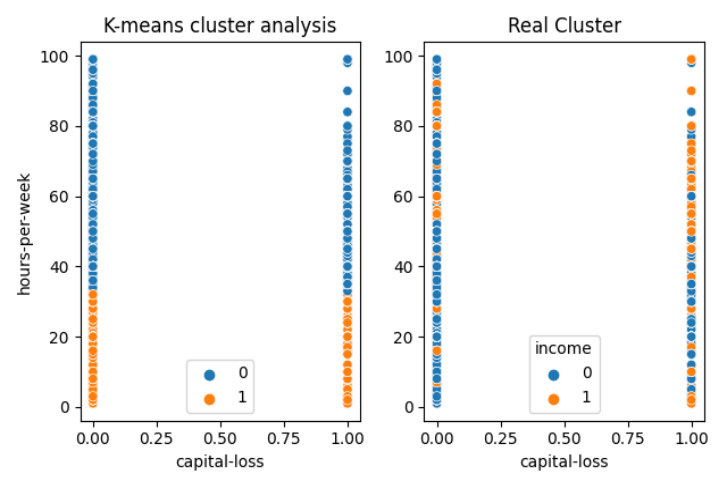
Age vs. capital gain

K-MEANS ANALYSE

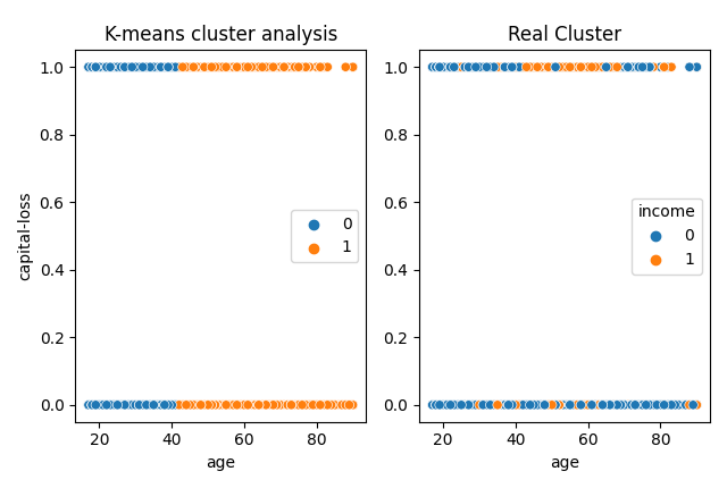
Capital gain vs. Capital loss



Capital loss vs. Hours per week

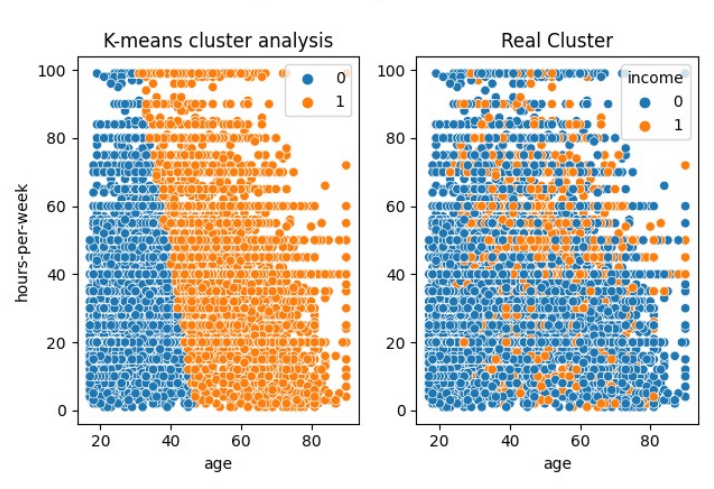


Age vs. Capital-Loss

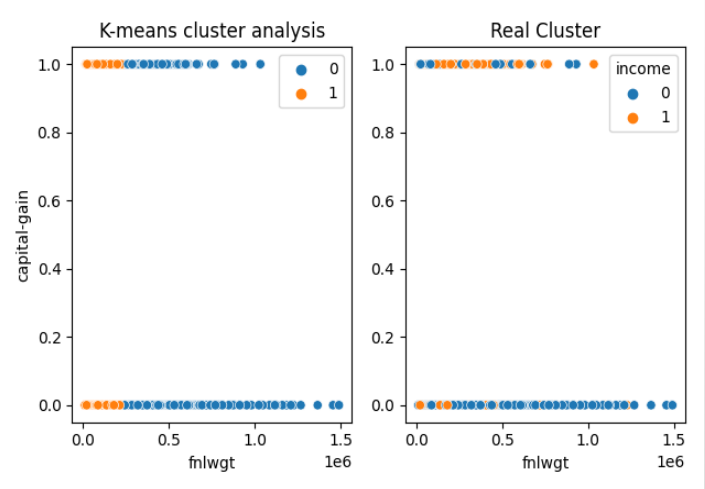


K-MEANS ANALYSE

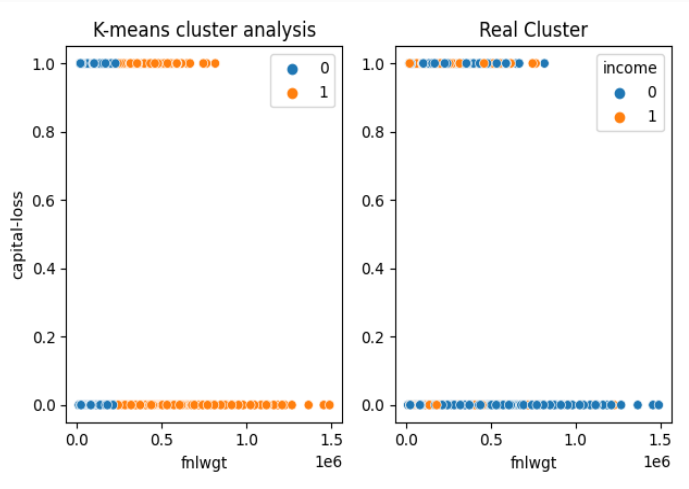
Age vs Hours-per-week



Financial weight vs Capital-gain

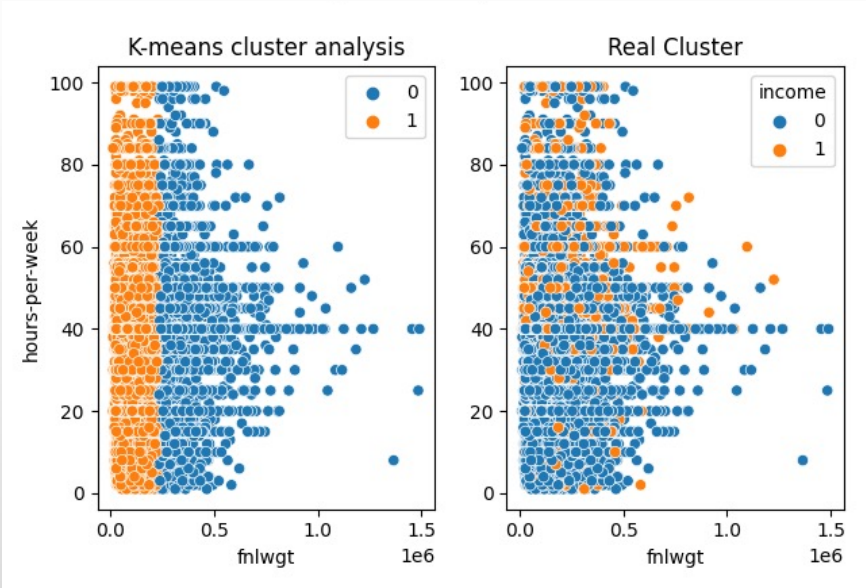


Financial weight vs Capital-loss

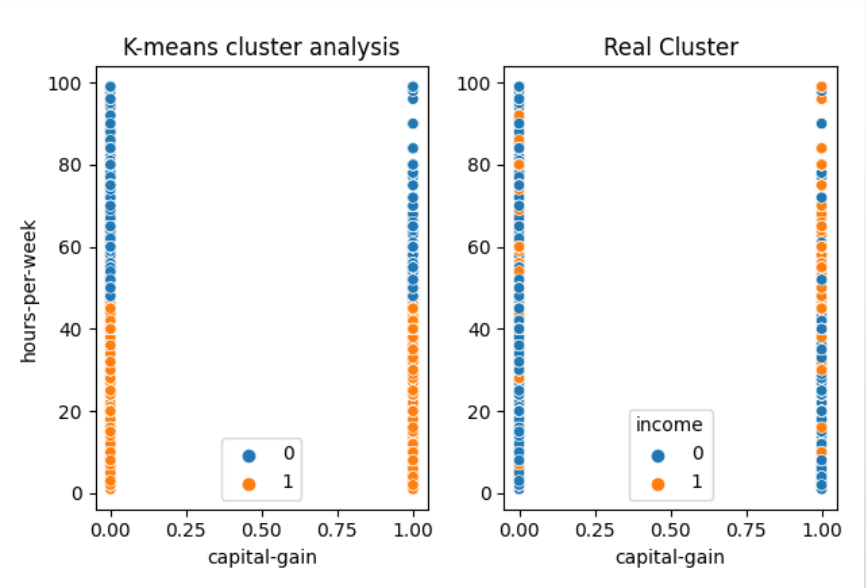


K-MEANS ANALYSE

Financial weight vs Hours-per-week



Capital-gain vs Hours-per-week



LOGISTISCHE KLASSIFIKATION

LOGIT MODELL

Einfache lineare Daten (skaliert)

- Relationship 4 hat am wenigsten Einfluss
- Working class 1 is am wenigsten signifikant in diesem Modell
- Education 1 & Capital-gain sind am signifikantesten

Results: Logit

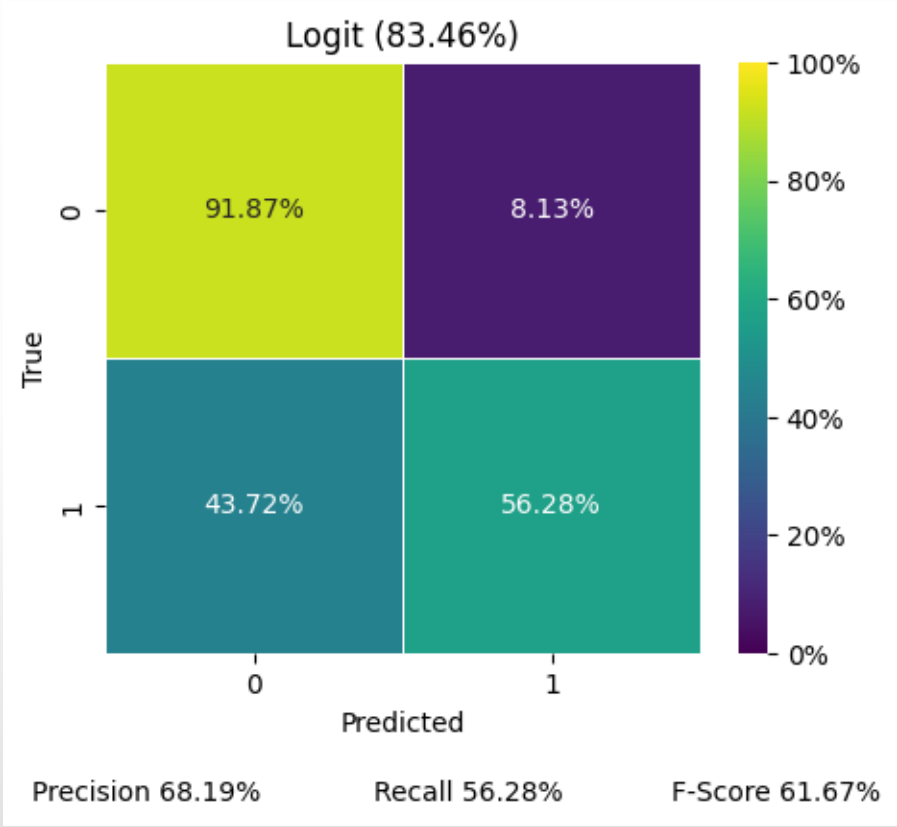
Model:	Logit	Method:	MLE
Dependent Variable:	income	Pseudo R-squared:	0.354
Date:	2023-06-21 14:51	AIC:	27843.0833
No. Observations:	39048	BIC:	27988.8166
Df Model:	16	Log-Likelihood:	-13905.
Df Residuals:	39031	LL-Null:	-21525.
Converged:	1.0000	LLR p-value:	0.0000
No. Iterations:	8.0000	Scale:	1.0000

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-6.2634	0.2039	-30.7245	0.0000	-6.6630	-5.8639
age	1.8523	0.0946	19.5897	0.0000	1.6670	2.0376
workclass_1	-0.1456	0.0788	-1.8478	0.0646	-0.3001	0.0088
workclass_2	-0.7941	0.0838	-9.4740	0.0000	-0.9583	-0.6298
workclass_3	-0.3948	0.0716	-5.5171	0.0000	-0.5351	-0.2546
fnlwgt	0.8810	0.2132	4.1328	0.0000	0.4632	1.2989
education_1	2.1186	0.0702	30.1894	0.0000	1.9811	2.2561
education_2	1.1905	0.0719	16.5458	0.0000	1.0494	1.3315
occupation_1	-0.6529	0.0326	-20.0273	0.0000	-0.7168	-0.5890
relationship_1	0.4364	0.1631	2.6752	0.0075	0.1167	0.7561
relationship_2	2.5018	0.1609	15.5502	0.0000	2.1865	2.8172
relationship_3	-0.8872	0.1938	-4.5789	0.0000	-1.2669	-0.5074
relationship_4	0.0891	0.1736	0.5131	0.6079	-0.2512	0.4293
sex_1	0.1681	0.0425	3.9518	0.0001	0.0847	0.2514
capital-gain	1.7106	0.0485	35.2427	0.0000	1.6154	1.8057
capital-loss	1.2101	0.0620	19.5328	0.0000	1.0887	1.3316
hours-per-week	3.2745	0.1302	25.1531	0.0000	3.0194	3.5297

LOGIT MODELL

Confusion Matrix

- 83,46% accuracy
- Einkommen unter 50k $ (0) wird wesentlich besser erkannt



LOGIT MODELL

Mit 8 Hauptkomponenten

- Enorm hohe Signifikanz aller Kovariablen (bis auf die 7. Hauptkomponente)
  - Liegt an der PCA

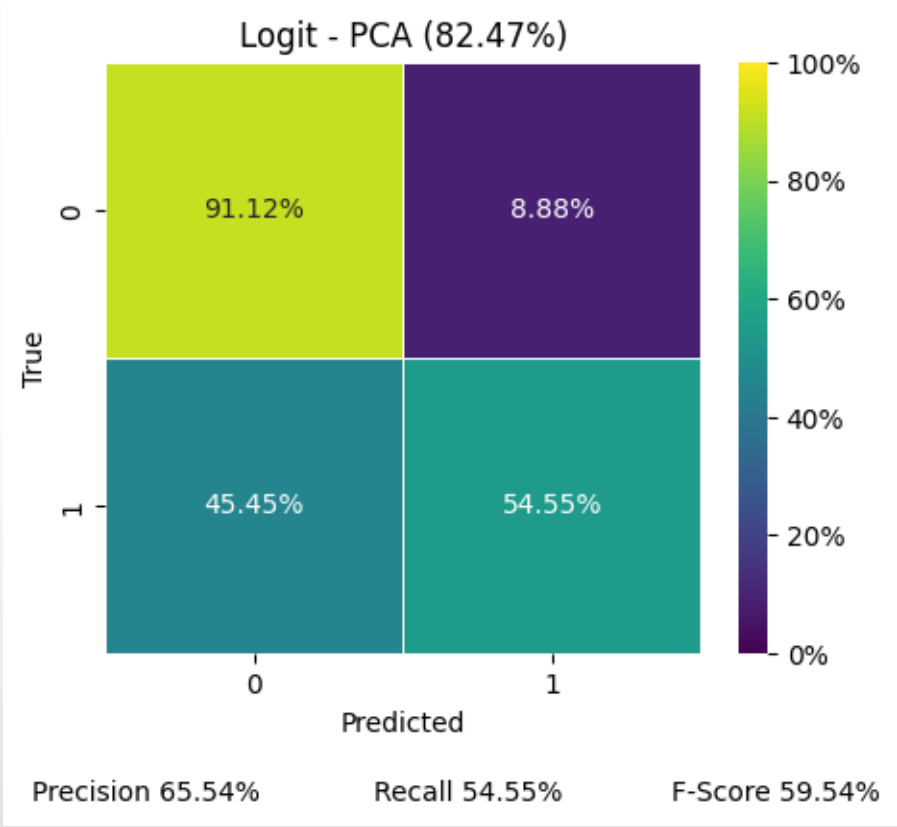
Model:			Logit	Method:		MLE
Dependent Variable:			income	Pseudo R-squared:		0.296
Date:			2023-06-21 14:51	AIC:		30334.5907
No. Observations:			39048	BIC:		30411.7436
Df Model:			8	Log-Likelihood:		-15158.
Df Residuals:			39039	LL-Null:		-21525.
Converged:			1.0000	LLR p-value:		0.0000
No. Iterations:			10.0000	Scale:		1.0000
	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-2.1603	0.0401	-53.9003	0.0000	-2.2388	-2.0817
0	-1.0055	0.0225	-44.7392	0.0000	-1.0496	-0.9615
1	2.4421	0.0454	53.8488	0.0000	2.3532	2.5310
2	0.3600	0.0447	8.0539	0.0000	0.2724	0.4476
3	0.2265	0.0380	5.9570	0.0000	0.1520	0.3010
4	-2.5083	0.1087	-23.0796	0.0000	-2.7213	-2.2953
5	-2.7608	0.1453	-19.0054	0.0000	-3.0455	-2.4761
6	0.1169	0.0783	1.4925	0.1356	-0.0366	0.2704
7	-0.5517	0.0494	-11.1791	0.0000	-0.6485	-0.4550

LOGIT MODELL

Confusion Matrix

- 82,47% accuracy
- Einkommen unter 50k $ (0) wird wesentlich besser erkannt
- Einkommen über 50k $ (1) wird schlechter erkannt

➢ Ähnliche Performance trotz weniger Variablen



LOGIT MODELL

Mit nicht linear transformierten Daten

- Working class 1 & relationship 4 sind weniger signifikant
  - Working class 1 hat ein breites KI trotz kleines Standardfehlers
- Education 1 & capital Gain sind sehr signifikant

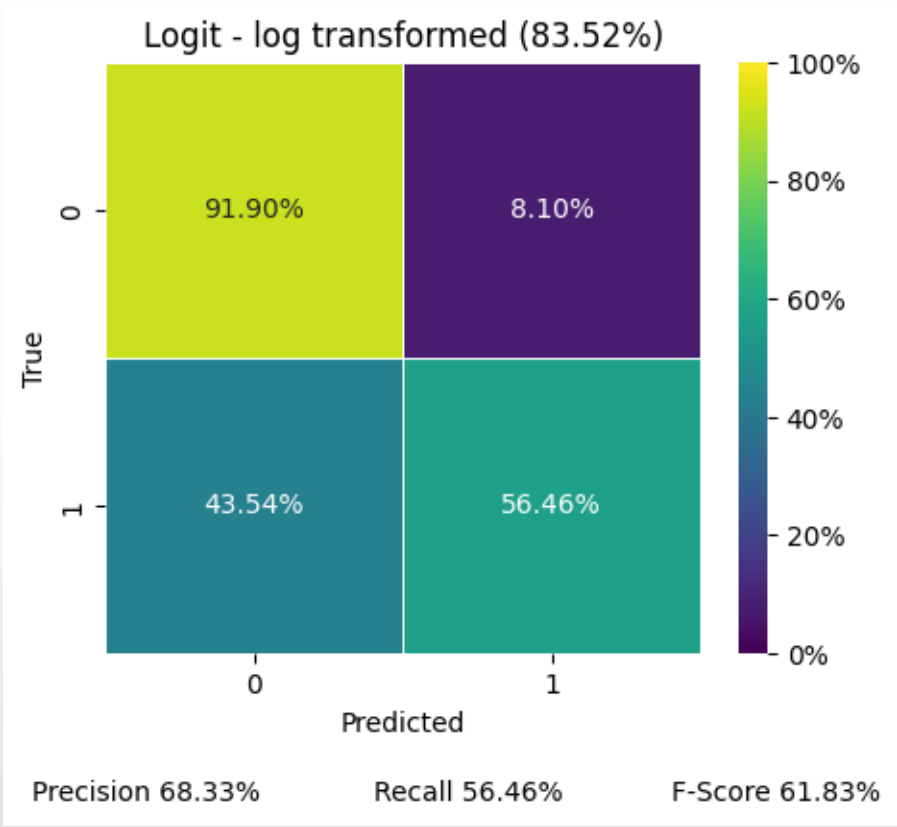
Model:	Logit			Method:		MLE
Dependent Variable:	income			Pseudo R-squared:		0.356
Date:	2023-06-21 14:51			AIC:		27760.2551
No. Observations:	39048			BIC:		27905.9884
Df Model:	16			Log-Likelihood:		-13863.
Df Residuals:	39031			LL-Null:		-21525.
Converged:	1.0000			LLR p-value:		0.0000
No. Iterations:	8.0000			Scale:		1.0000
	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-6.4409	0.2056	-31.3230	0.0000	-6.8439	-6.0379
age	2.8015	0.1305	21.4622	0.0000	2.5457	3.0574
workclass_1	-0.1421	0.0789	-1.8015	0.0716	-0.2968	0.0125
workclass_2	-0.7931	0.0839	-9.4523	0.0000	-0.9575	-0.6286
workclass_3	-0.3812	0.0716	-5.3200	0.0000	-0.5216	-0.2407
fnlwgt	1.0927	0.2475	4.4148	0.0000	0.6076	1.5778
education_1	2.1283	0.0702	30.3165	0.0000	1.9907	2.2659
education_2	1.1963	0.0720	16.6198	0.0000	1.0552	1.3374
occupation_1	-0.6502	0.0326	-19.9238	0.0000	-0.7141	-0.5862
relationship_1	0.4195	0.1635	2.5662	0.0103	0.0991	0.7398
relationship_2	2.4710	0.1612	15.3250	0.0000	2.1550	2.7870
relationship_3	-0.8452	0.1941	-4.3535	0.0000	-1.2257	-0.4647
relationship_4	0.0544	0.1739	0.3130	0.7543	-0.2865	0.3954
sex_1	0.1651	0.0426	3.8741	0.0001	0.0816	0.2486
capital-gain	1.7066	0.0486	35.1159	0.0000	1.6113	1.8018
capital-loss	1.2066	0.0620	19.4521	0.0000	1.0850	1.3282
hours-per-week	3.2762	0.1301	25.1806	0.0000	3.0212	3.5312

LOGIT MODELL

Confusion Matrix

- Minimale Verbesserung zum Ursprungsmodell

➢ Logarithmische Transformation auf den 2 Variablen bringt nicht wirklich viel

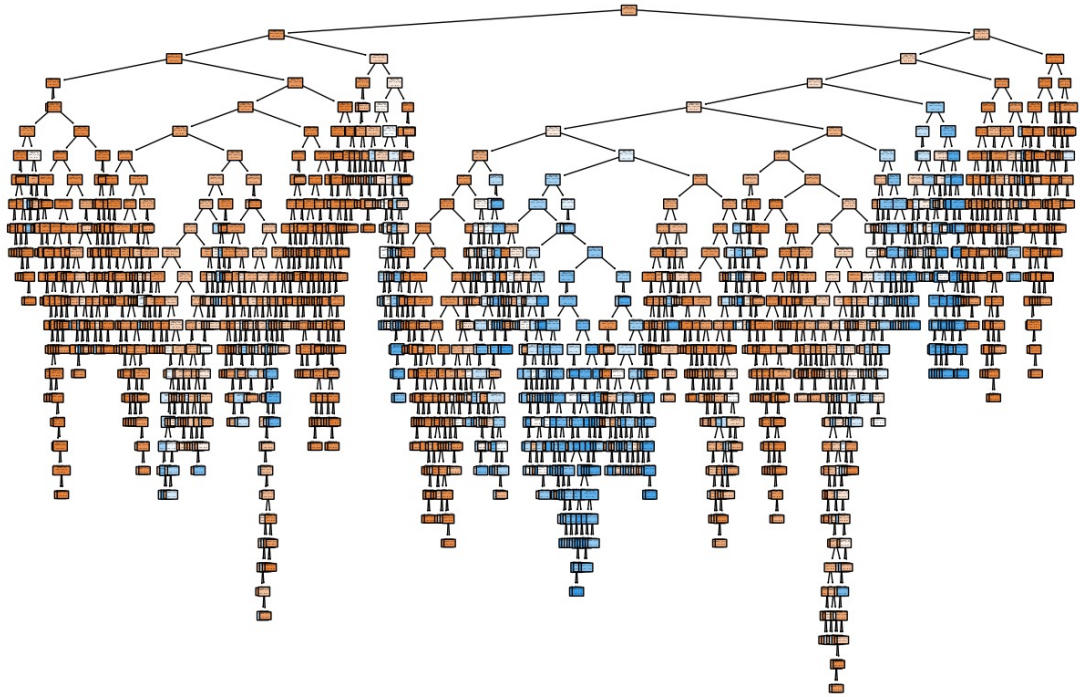


RANDOM FOREST

RANDOM FOREST

Decision Tree

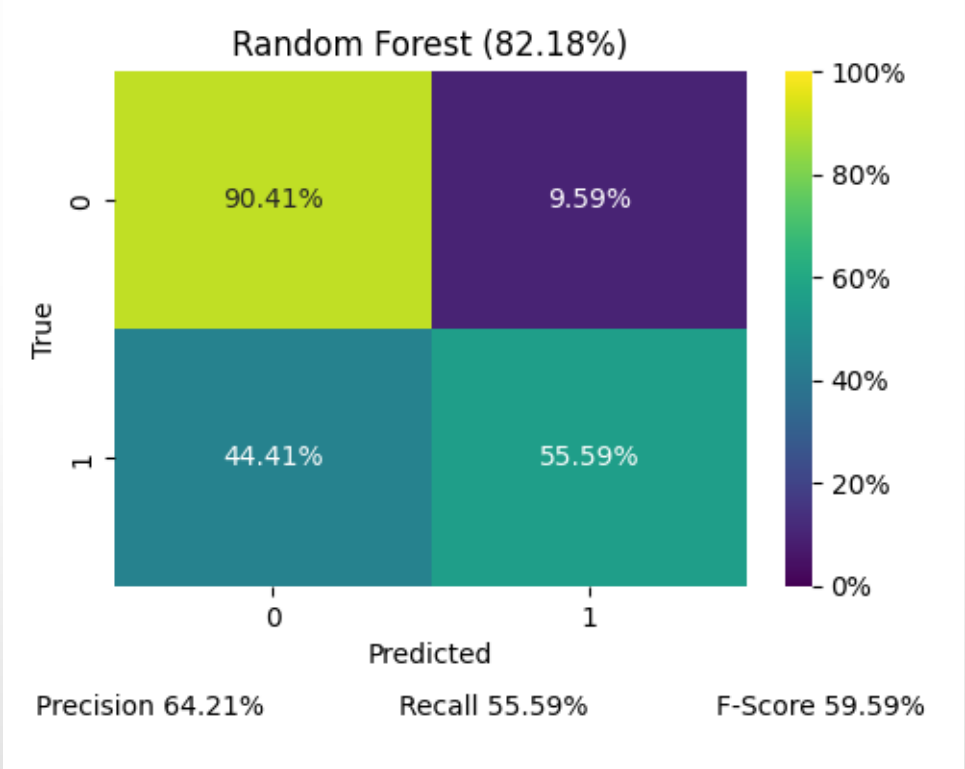
- Einer der Bäume des RF



RANDOM FOREST

Confusion Matrix

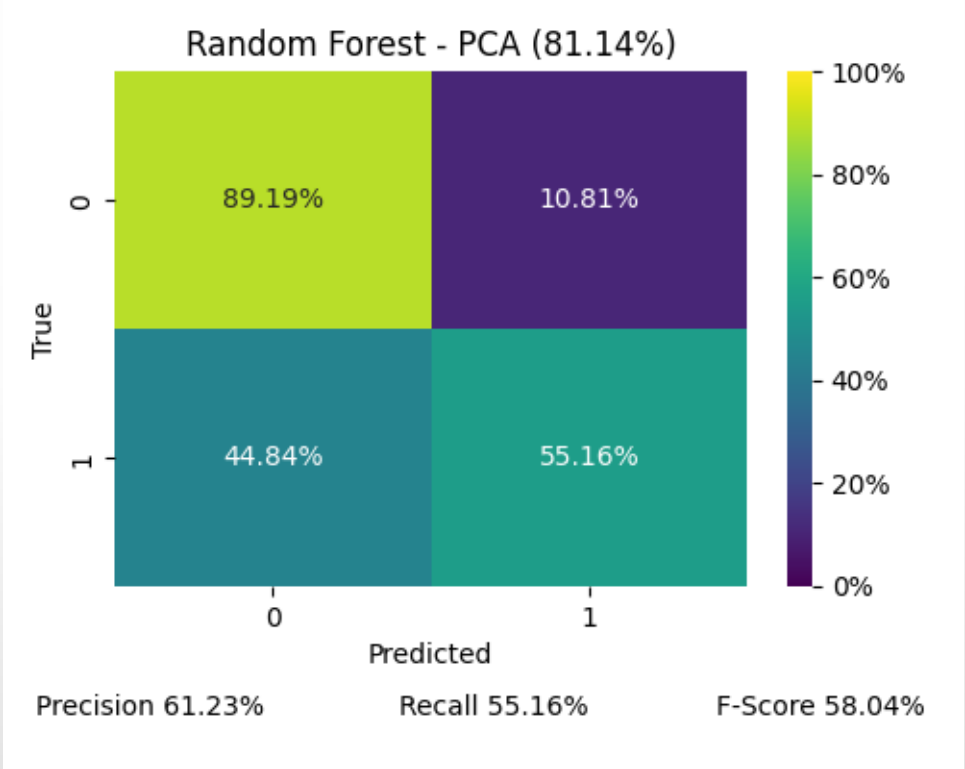
- 82,18% accuracy
- Verschlechterung zum LOGIT Modell



RANDOM FOREST

Confusion Matrix

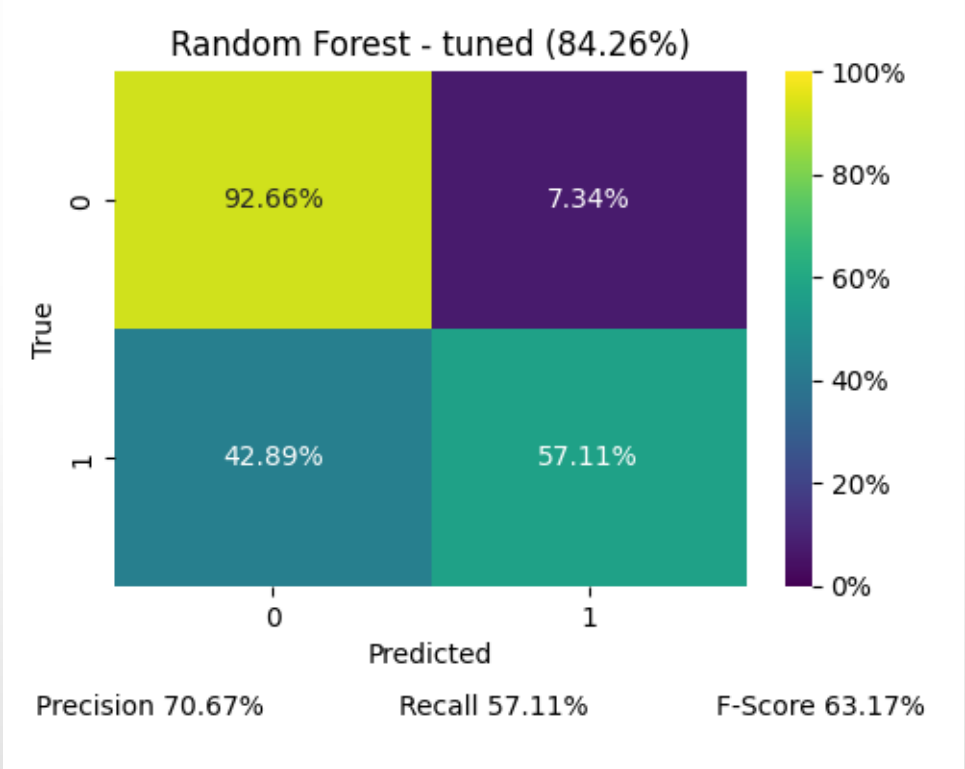
- Hat wohl nix gebracht.
  - Schlechtere Performance



RANDOM FOREST

Confusion Matrix

- Beste Performance der zufälligen Wälder

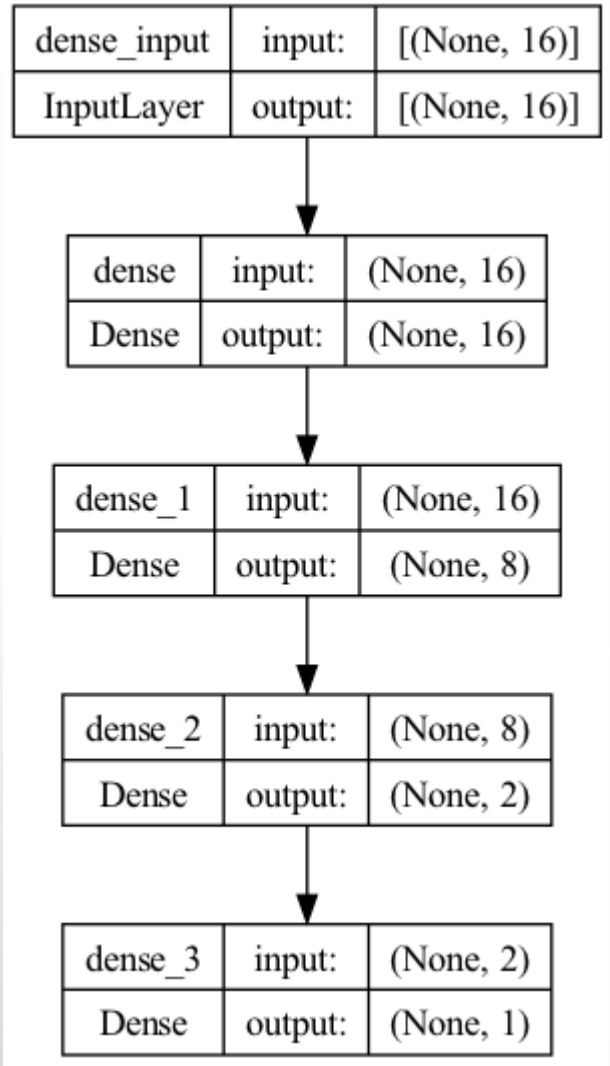


NEURONALES NETZ

NEURONALES NETZ

Erste Gehversuche

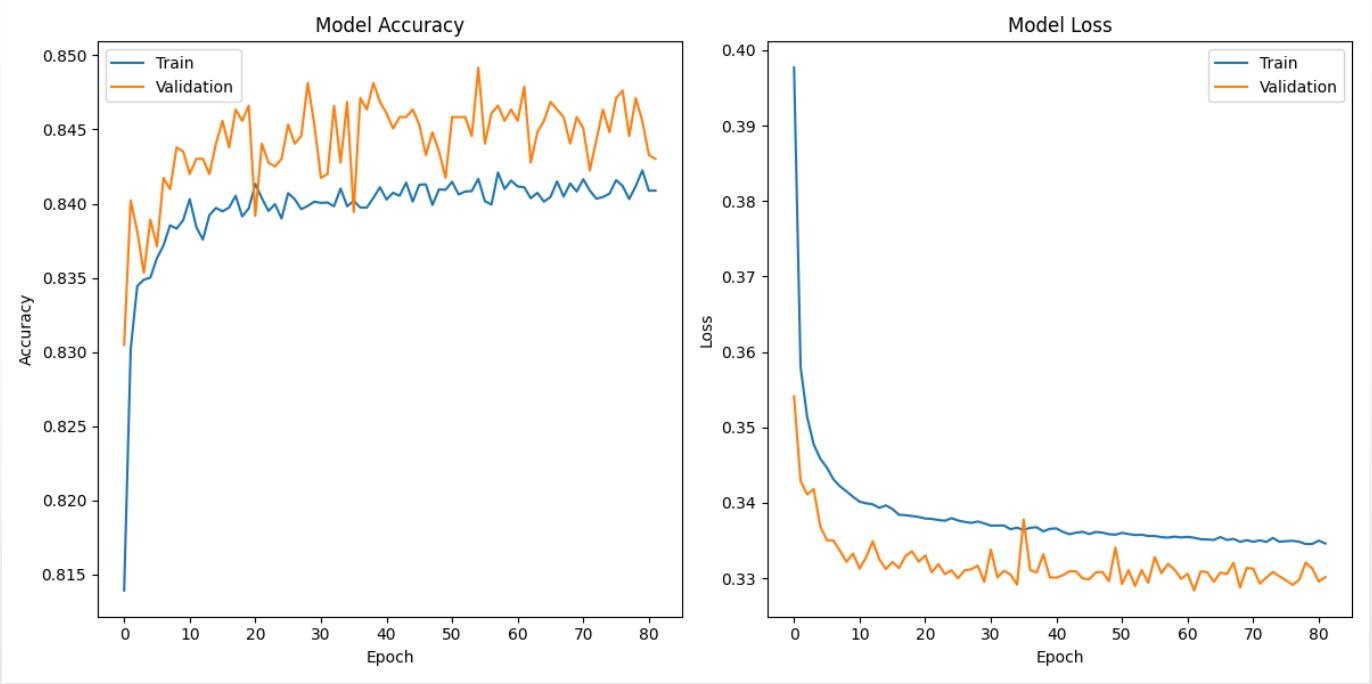
- Inputschicht
  - Anzahl Neuronen = Anzahl an Features
- 4 Dense Schichten
- Output: binär



NEURONALES NETZ

Training & Validierung

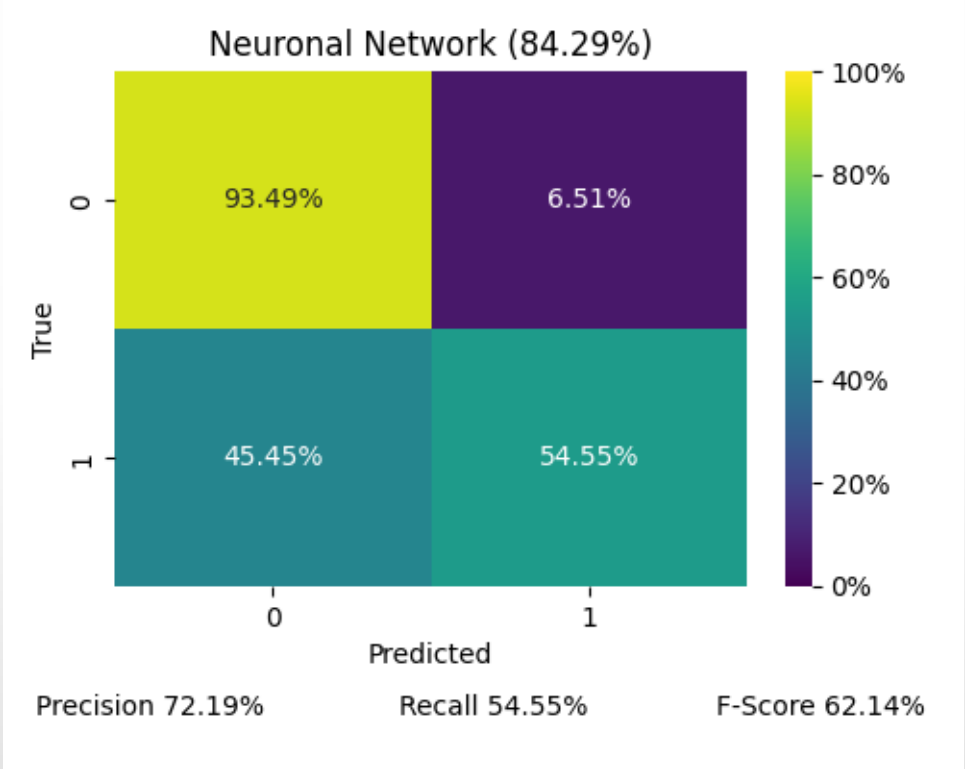
- Early stopping nach 20 perioden
- Loss Funktion: Binary crossentropy
- Validationsplit: 10%
- Batch size: 16



NEURONALES NETZ

Confusion Matrix

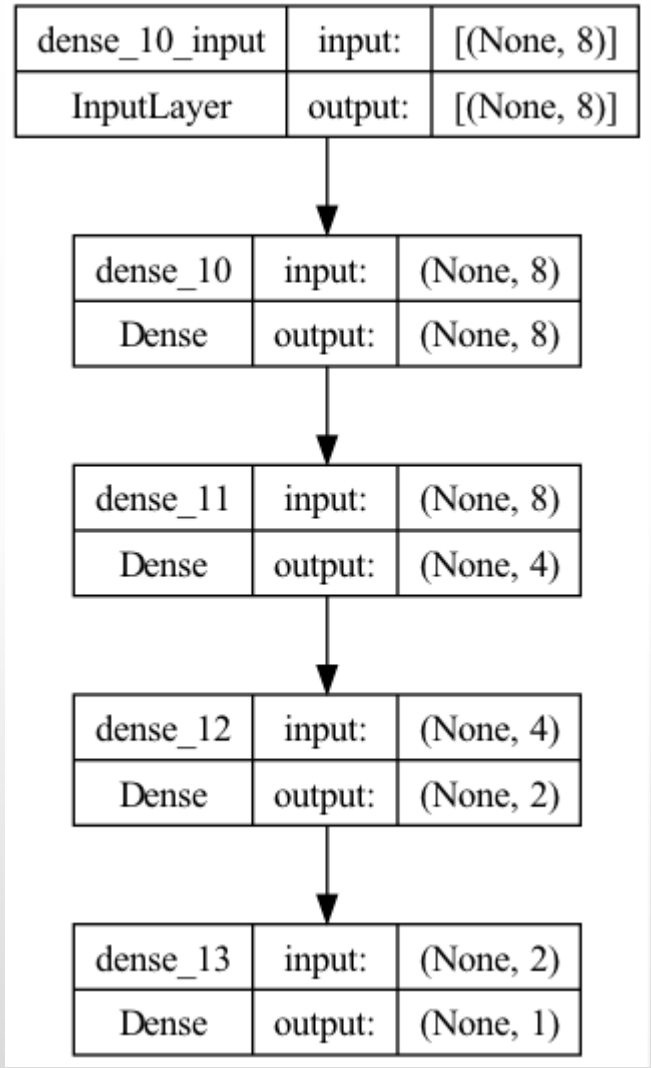
- 84,29% accuracy
- Einkommen über 50k $ wird bisher am besten vorhergesagt



NEURONALES NETZ

Zweiter Gehversuch

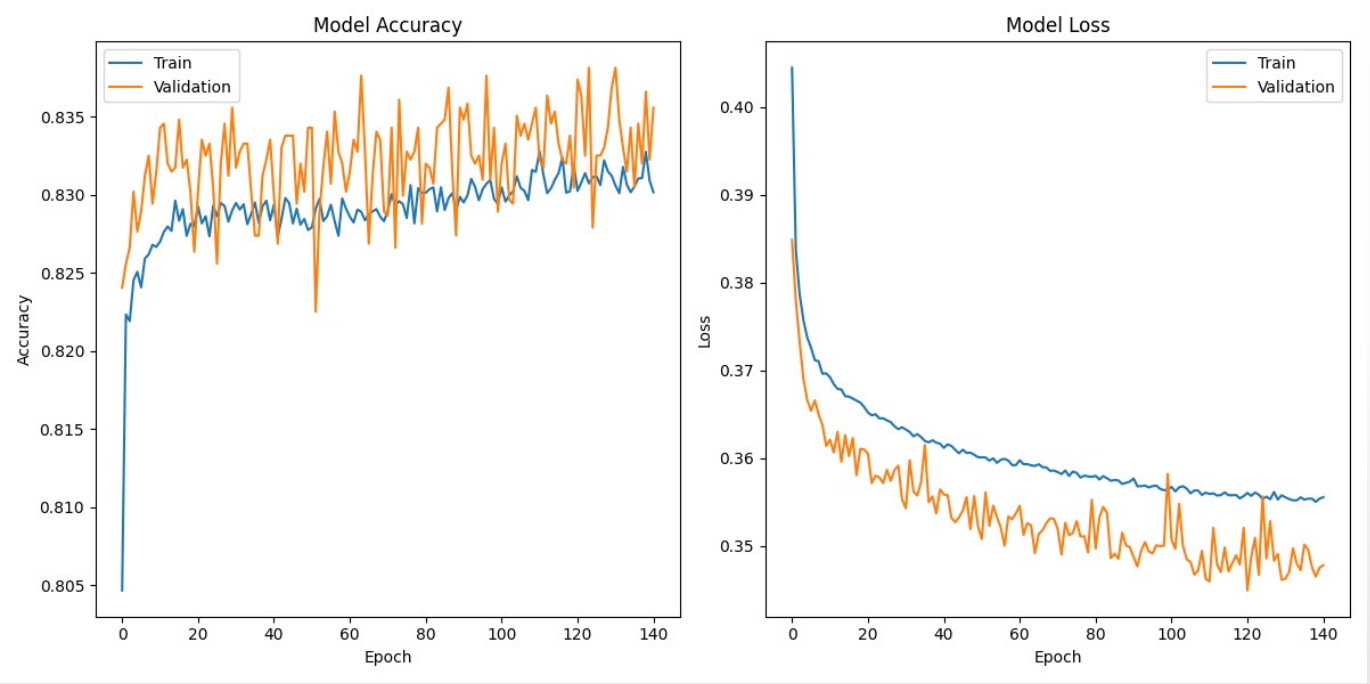
- Mit PCA Daten
- Inputschicht
  - Anzahl Neuronen = Anzahl an Hauptkomponenten
- 4 Dense Schichten
- Output: binär



NEURONALES NETZ

Training & Validierung

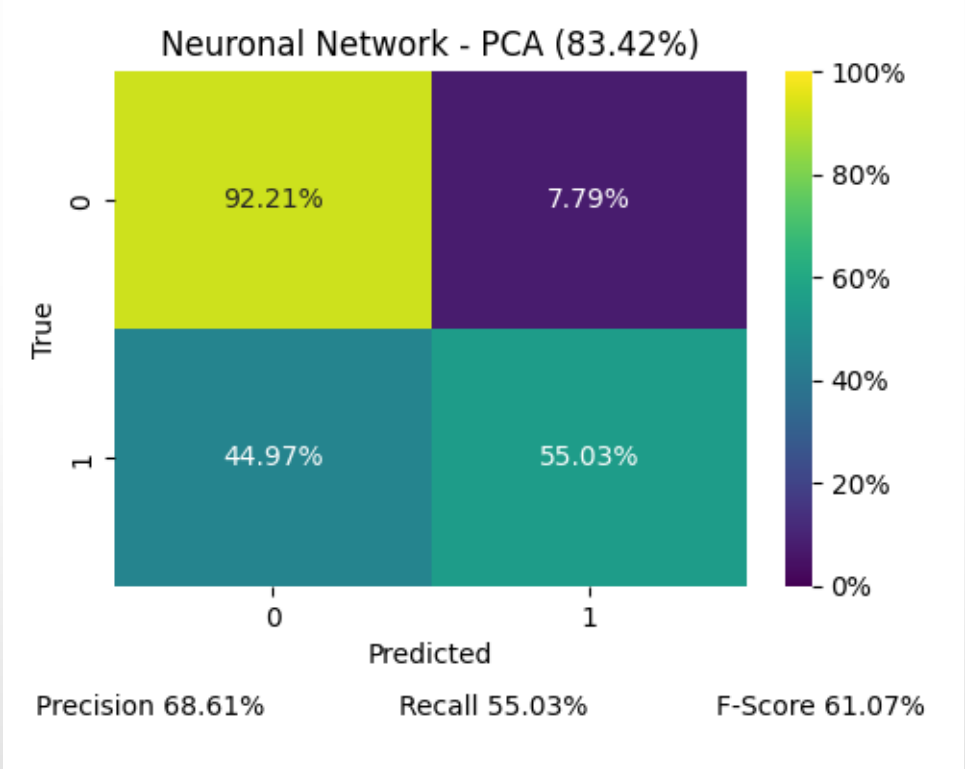
- Early stopping nach 20 perioden
- Loss Funktion: Binary crossentropy
- Validationsplit: 10%
- Batch size: 8



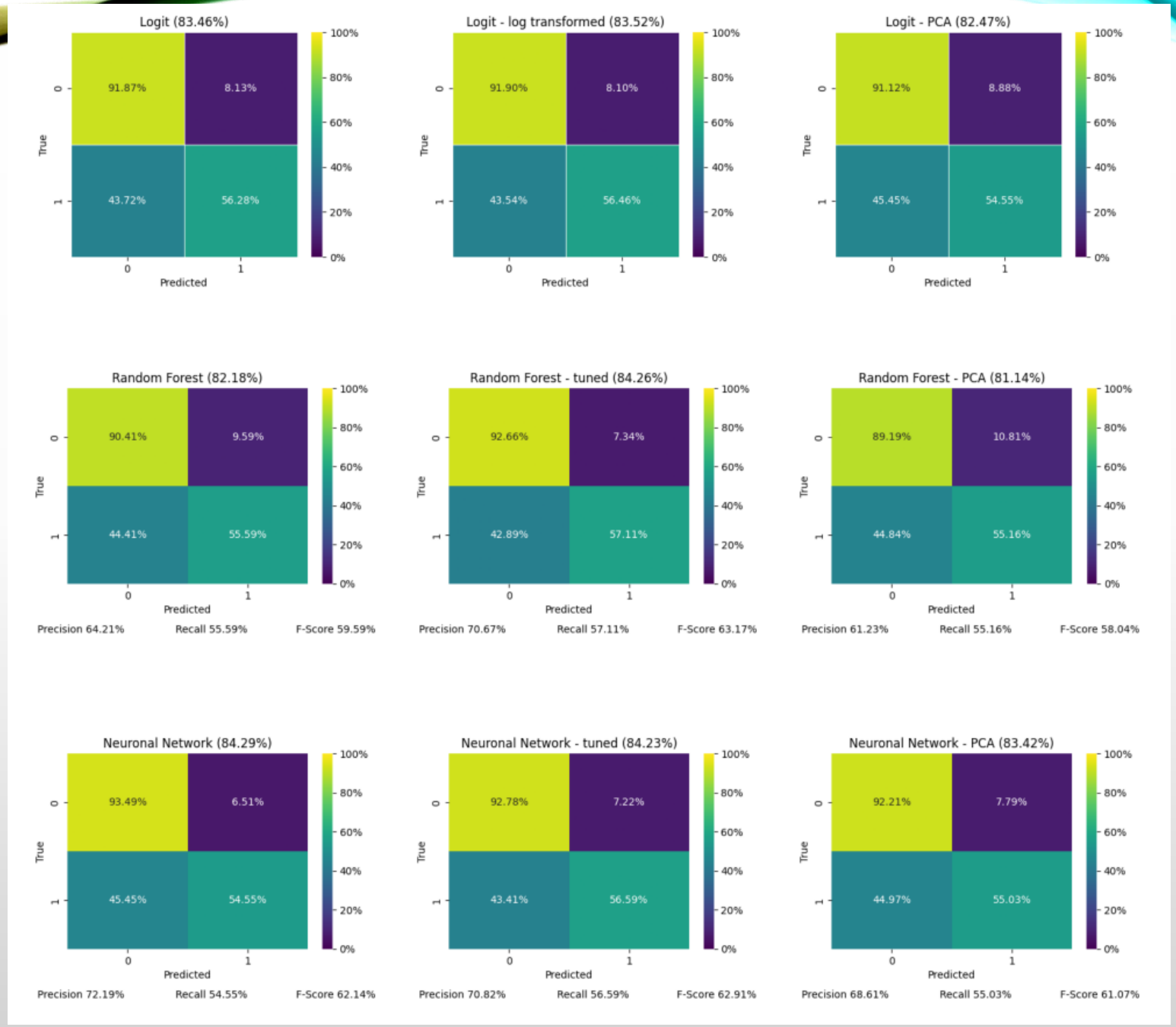
NEURONALES NETZ

Confusion Matrix

- Erneuter Rückgang bei der Vorhersage der Einkommen über 50k$



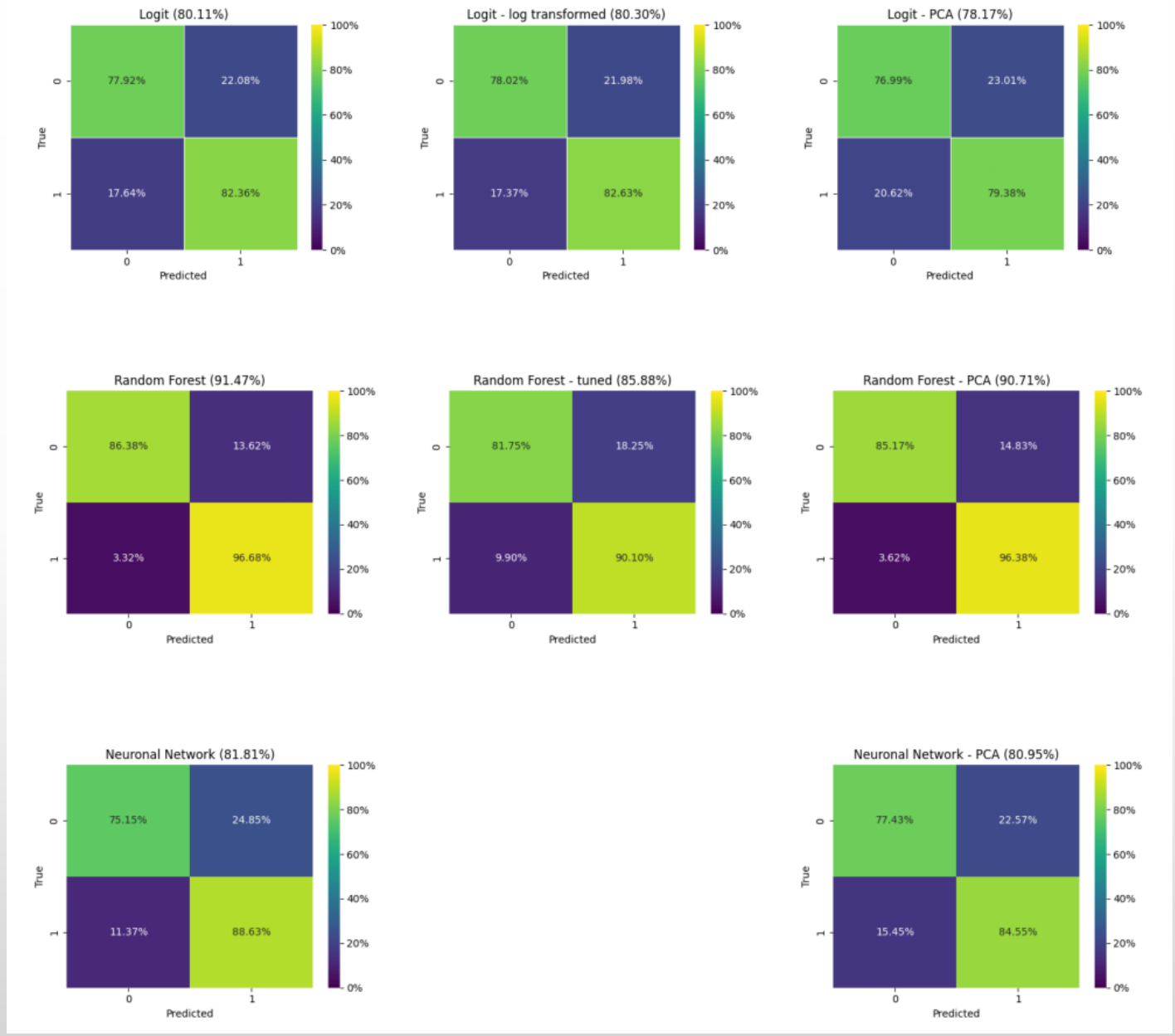
MODELL VERGLEICH



FRAGEN?

DANKE!

UPSAMPLING


Logit (80.11%)
77.92%
22.08%
17.64%
82.36%
Logit - log transformed (80.30%)
78.02%
21.98%
17.37%
82.63%
Logit - PCA (78.17%)
76.99%
23.01%
20.62%
79.38%
Random Forest (91.47%)
86.38%
13.62%
3.32%
96.68%
Random Forest - tuned (85.88%)
81.75%
18.25%
9.90%
90.10%
Random Forest - PCA (90.71%)
85.17%
14.83%
3.62%
96.38%
Neuronal Network (81.81%)
75.15%
24.85%
11.37%
88.63%
Neuronal Network - PCA (80.95%)
77.43%
22.57%
15.45%
84.55%
True
Predicted
0
1
100%
80%
60%
40%
20%
0%