Dataset «Stroke»

Shamakov Viktor

Shape of data train (20832, 12) Shape of data test (5208, 12)

1. Знакомимся с данными. Признаки

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	15784	Female	33.0	0	0	Yes	Private	Rural	79.53	23.1	never smoked	0
1	16049	Male	48.0	0	0	Yes	Govt_job	Rural	74.12	NaN	smokes	0
2	23616	Female	32.0	0	0	No	Private	Urban	69.83	22.0	never smoked	0
3	52824	Female	59.0	0	0	Yes	Private	Rural	146.97	30.1	never smoked	0
4	22322	Female	31.0	0	0	No	Private	Rural	69.27	28.9	NaN	0

- id - идентификатор пациента в наборе данных;
- gender - пол пациента;
- age - возраст пациента;
- hypertension - страдает ли пациент от гипертонии;
- heart_disease - страдает ли пациент от болезней сердца;
- ever_married - был ли пациент когда-либо женат;
- work_type - вид занятости;
- Residence_type - является ли пациент городским или сельским жителем;
- avg_glucose_level - средний уровень сахара в крови, который измерялся после еды;
- bmi - индекс массы тела;
- smoking_status - курит ли пациент;
- stroke - столбец правильных ответов: возникал ли у пациента инсульт.

1. Знакомимся с данными. Информация

Размер данных:

Shape of data train (20832, 12)

Shape of data test (5208, 12)

Пропуски:

bmi	681
smoking_status	6369

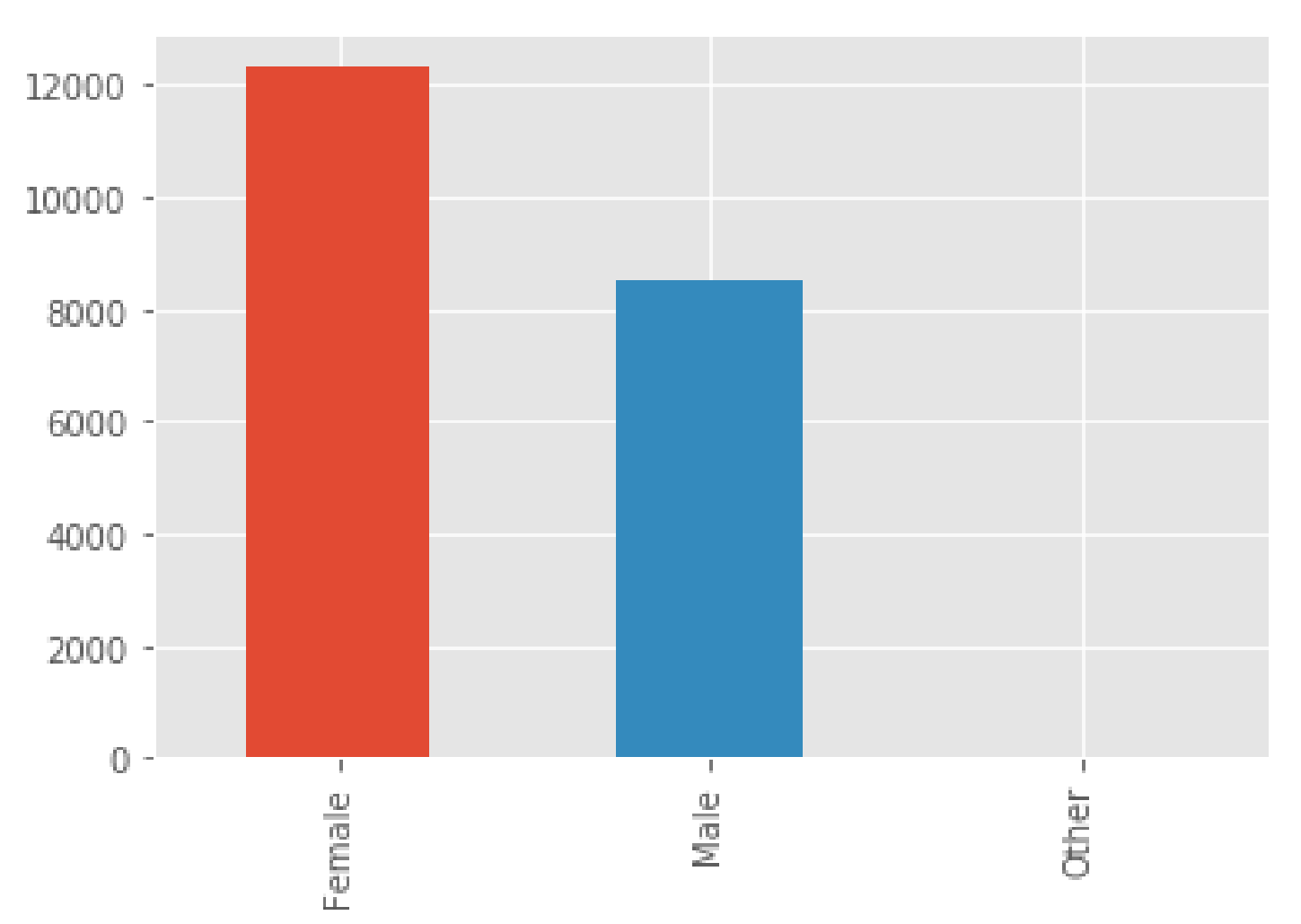
Типы данных:

id	int64
gender	object
age	float64
hypertension	int64
heart_disease	int64
ever_married	object
work_type	object
Residence_type	object
avg_glucose_level	float64
bmi	float64
smoking_status	object
stroke	int64

1. Знакомимся с данными. Описательные статистики

	age	hypertension	heart_disease	avg_glucose_level	bmi
count	20832.000000	20832.000000	20832.000000	20832.000000	20151.000000
mean	42.269677	0.093126	0.049395	104.357774	28.599687
std	22.562275	0.290616	0.216697	43.259818	7.783798
min	0.080000	0.000000	0.000000	55.000000	10.100000
25%	24.000000	0.000000	0.000000	77.390000	23.200000
50%	44.000000	0.000000	0.000000	91.320000	27.700000
75%	60.000000	0.000000	0.000000	111.470000	32.900000
max	82.000000	1.000000	1.000000	281.590000	96.100000

2.1. Подготовка данных. Gender



Female	12314
Male	8514
Other	4

2.1. Подготовка данных. Gender

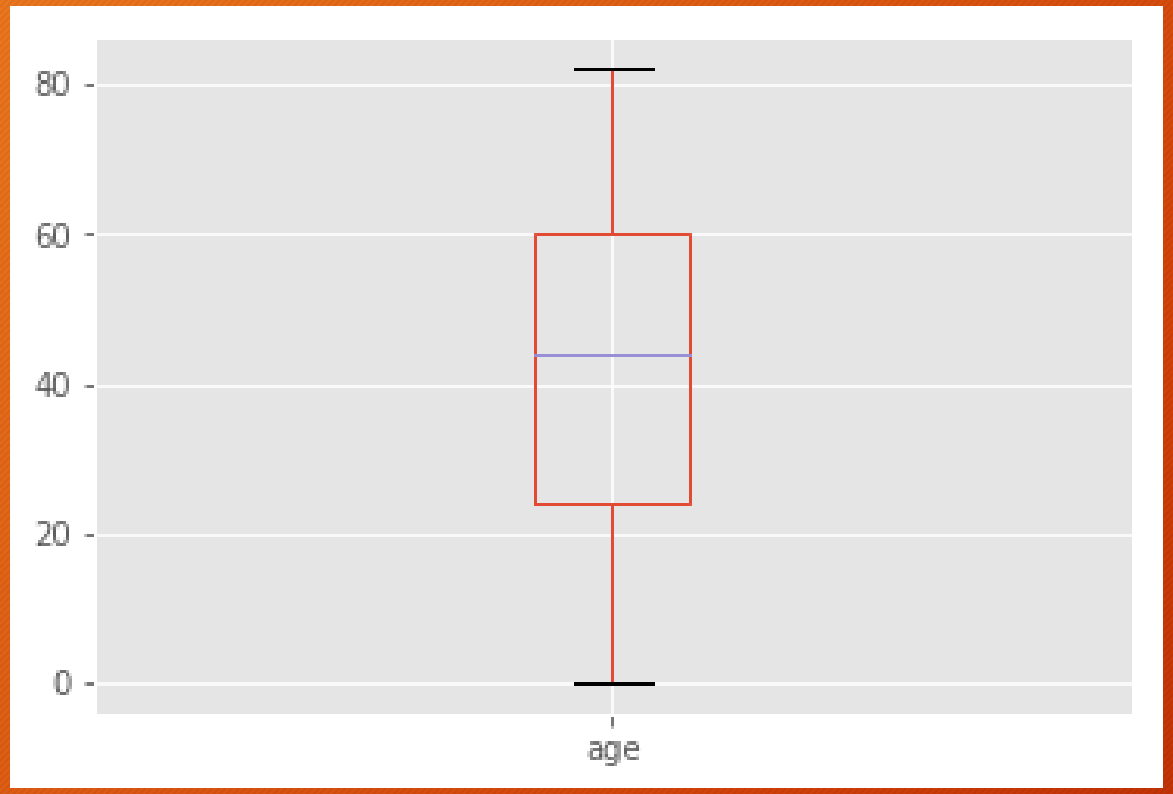

```
data_train.loc[data_train['gender'] == 'Other']
```

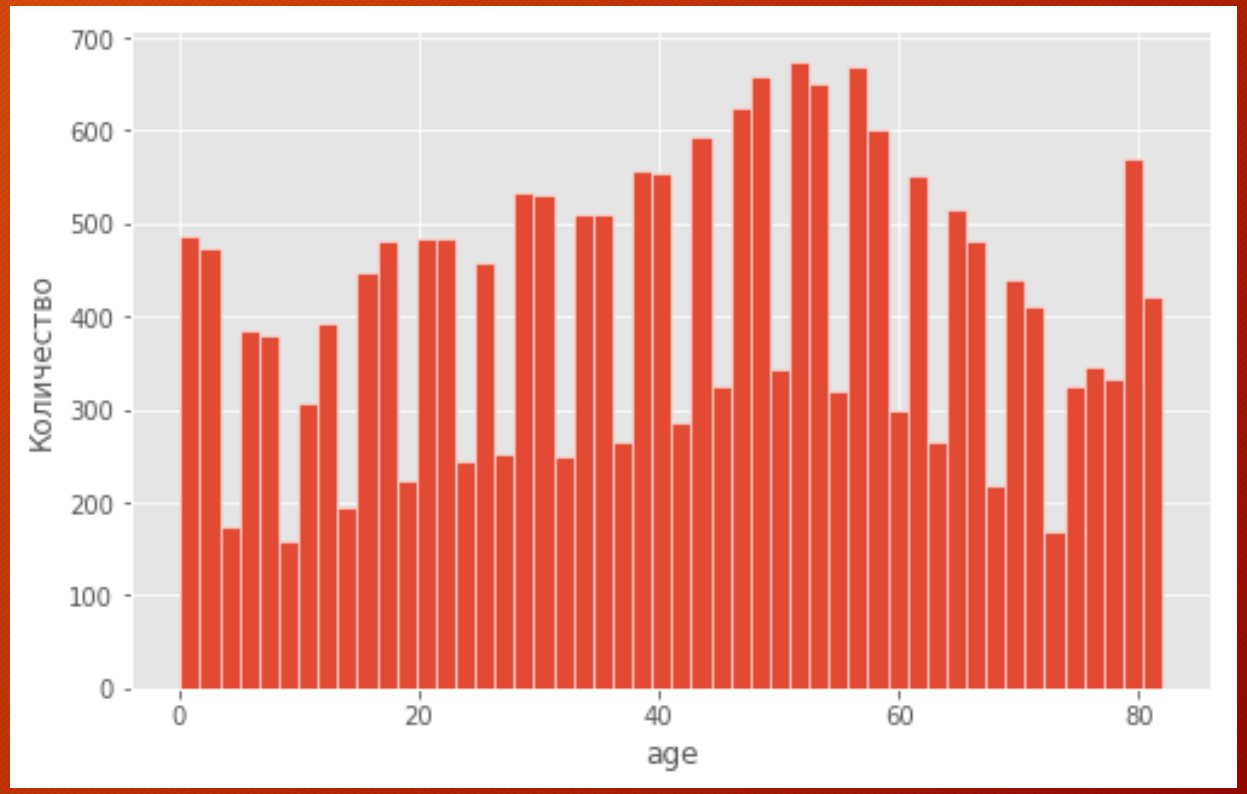
	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
2397	49812	Other	53.0	0	0	Yes	Govt_job	Urban	65.29	NaN	NaN	0
8984	42452	Other	10.0	0	0	No	children	Rural	72.05	21.2	formerly smoked	0
9320	13471	Other	39.0	0	0	Yes	Govt_job	Urban	95.59	32.0	formerly smoked	0
18411	26188	Other	46.0	0	0	No	Private	Rural	83.28	NaN	never smoked	0

Other - удаляем
Male меняем на 1
Female меняем на 0

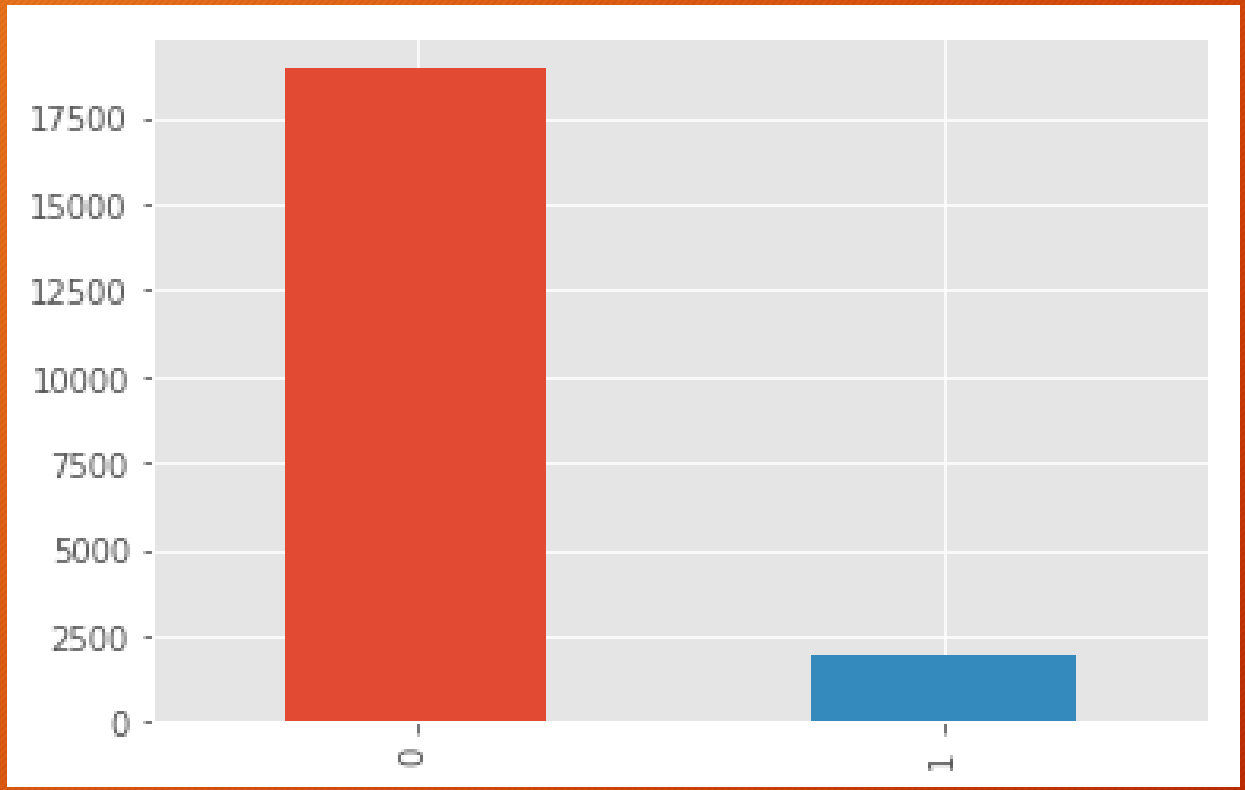
0 — 12314
1 — 8514

2.2. Подготовка данных. Age



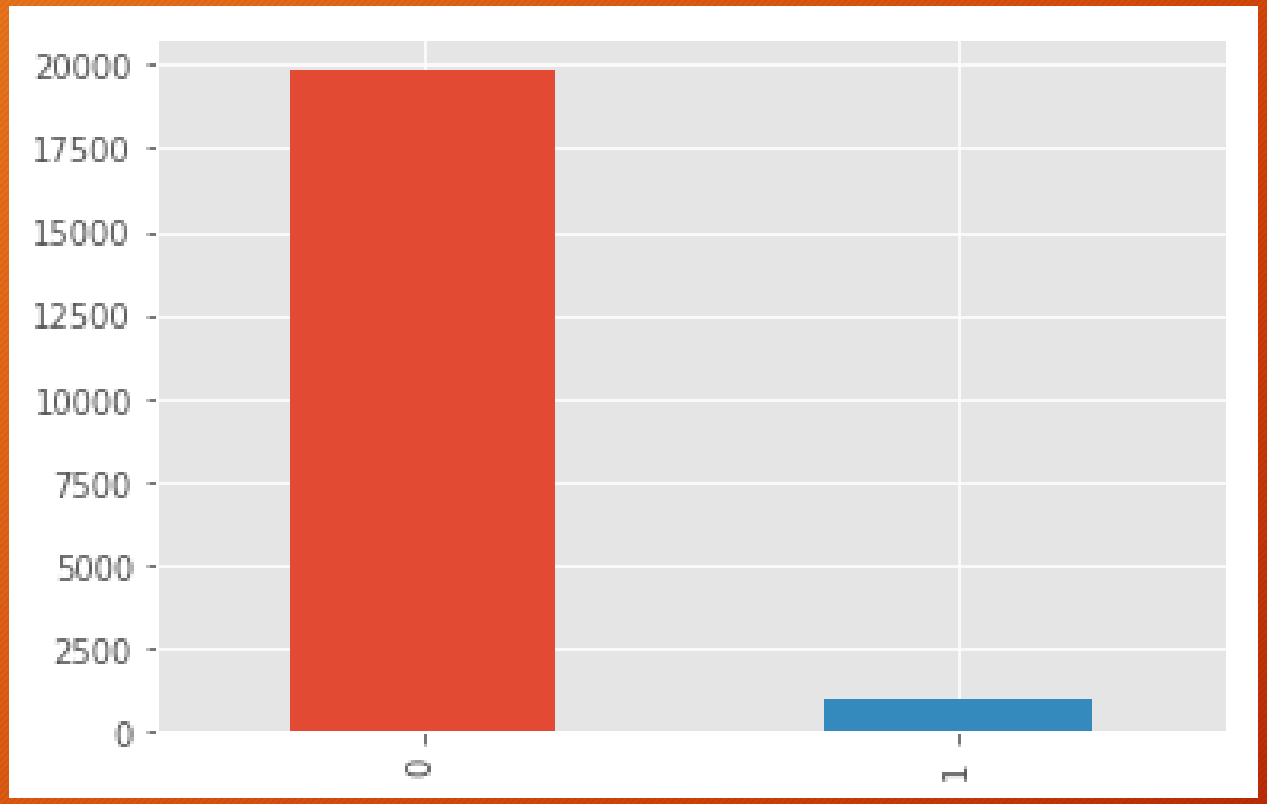


2.3. Подготовка данных. Hypertension



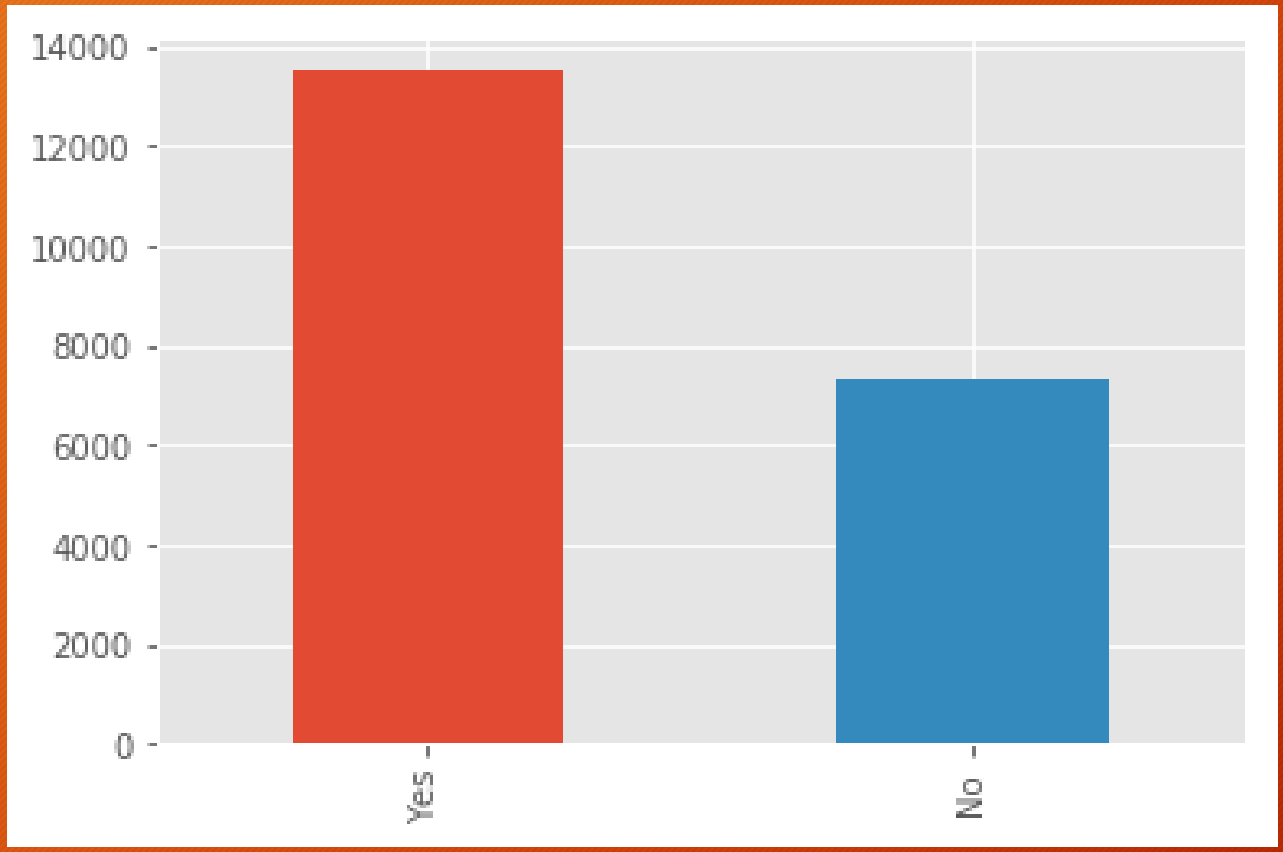
0 — 12314
1 — 8514

2.4. Подготовка данных. Heart Disease



0 – 19799
1 – 1029

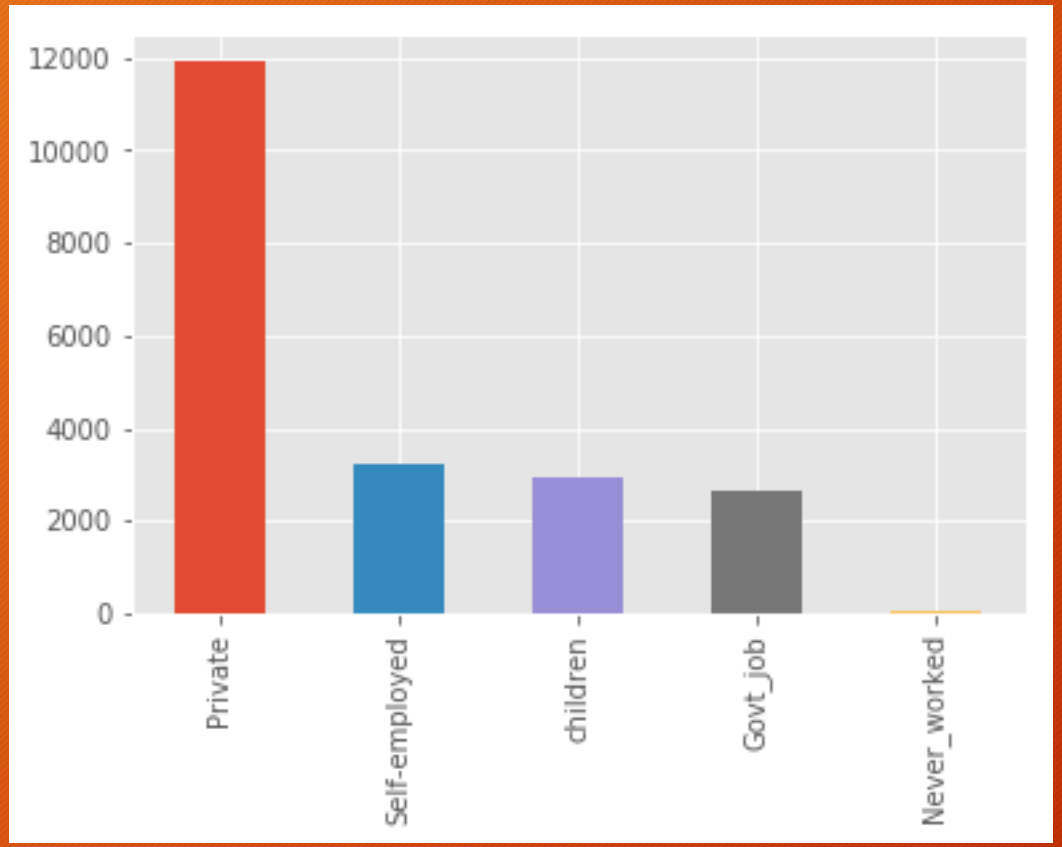
2.5. Подготовка данных. Ever married



Yes – 13503
No – 7325

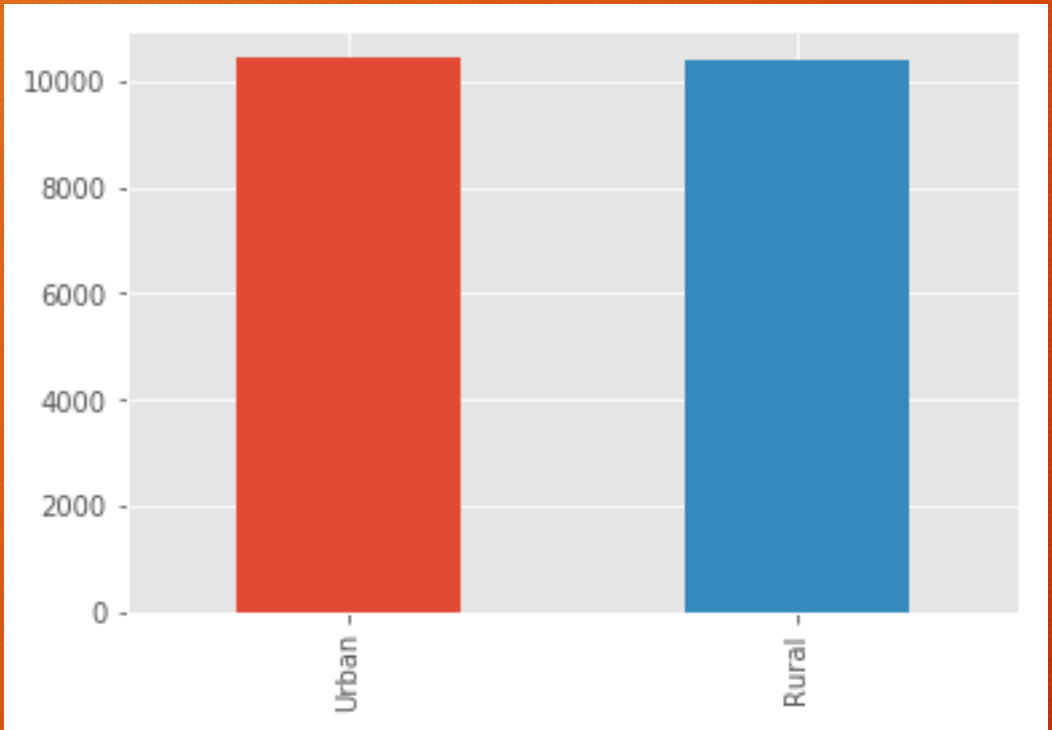
1 – 13503
0 – 7325

2.6. Подготовка данных. Work type



Private	11931
Self-employed	3229
children	2957
Govt_job	2632
Never_worked	79

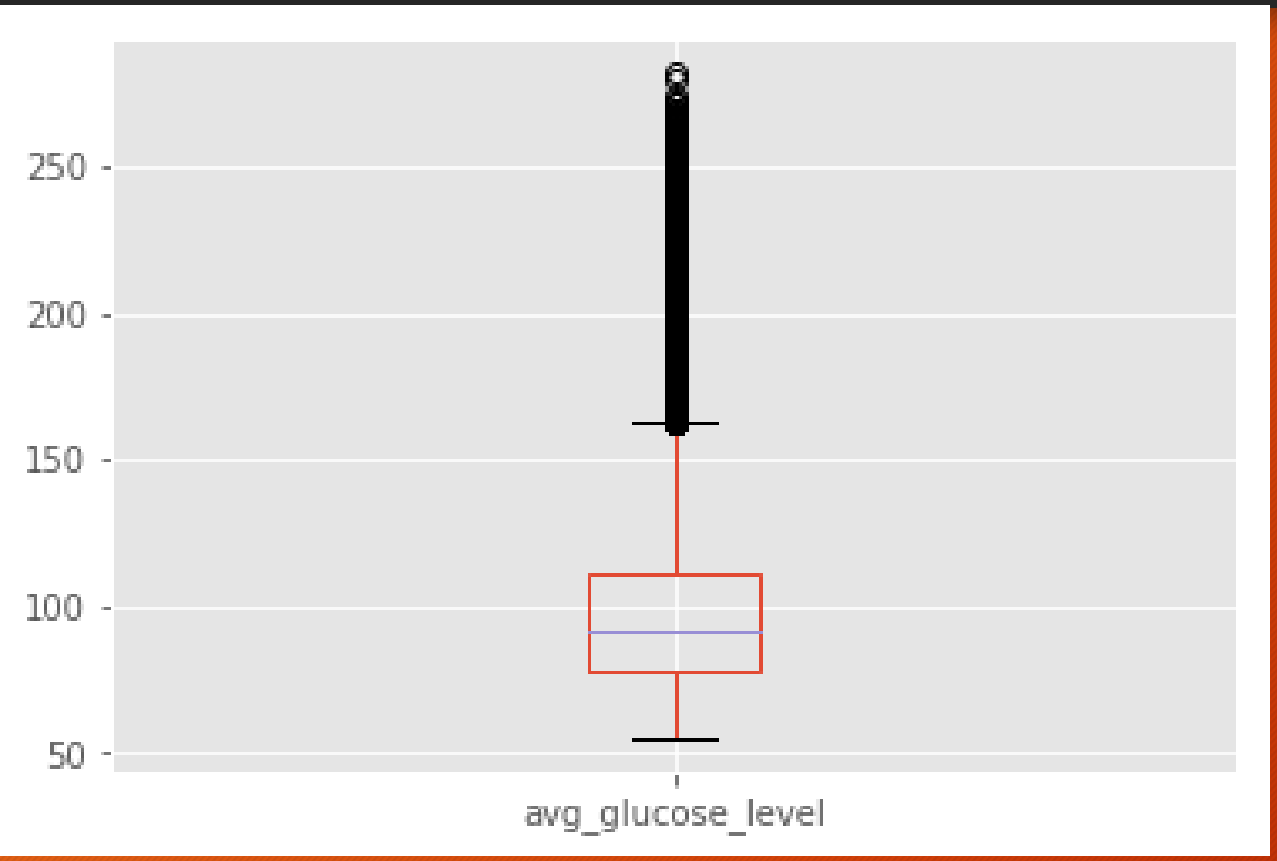
2.7. Подготовка данных. Residence type



Urban 10423
Rural 10405

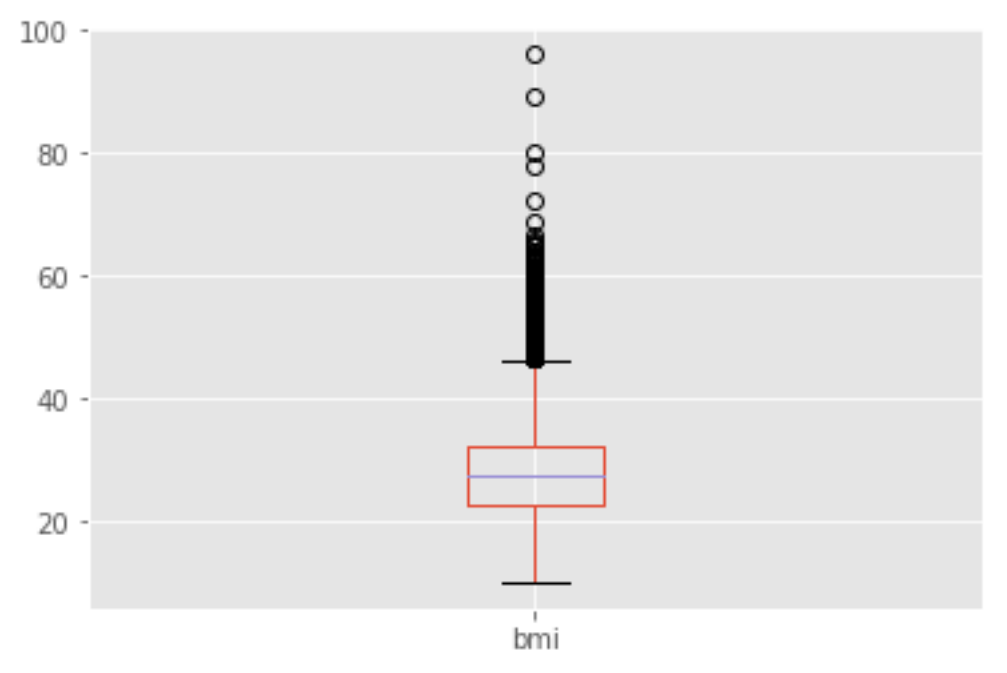
Заменяем на 0 и 1
1 10423
0 10405

2.8. Подготовка данных. Avg glucose level



Outliers пока оставляем без изменений
Пропусков нет

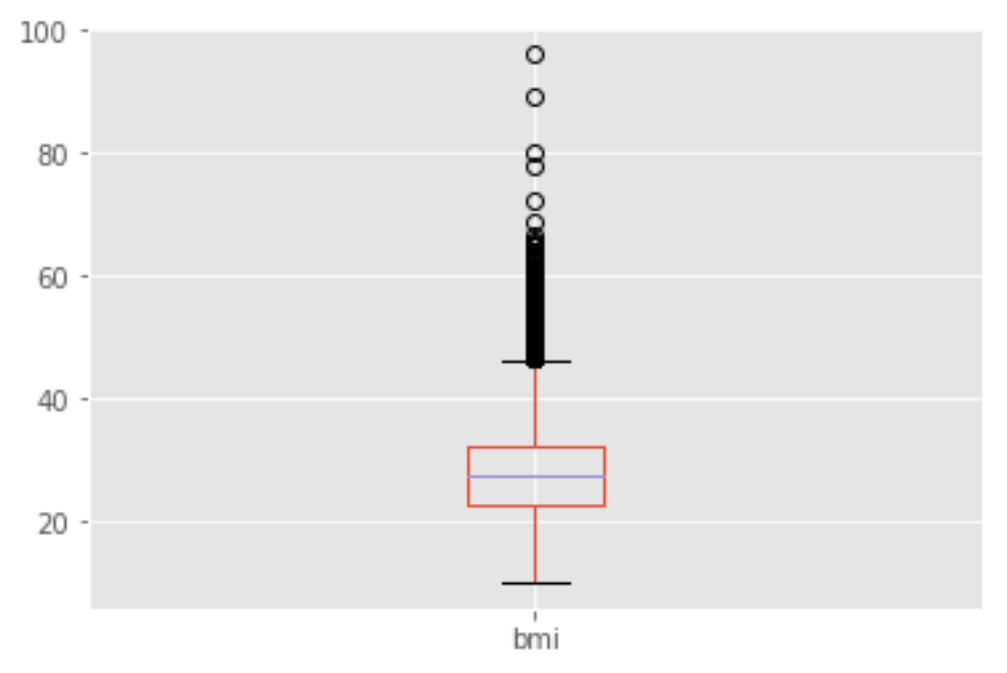
2.9. Подготовка данных. BMI



Outliers пока оставляем без изменений

Пропуски заполним средним значением

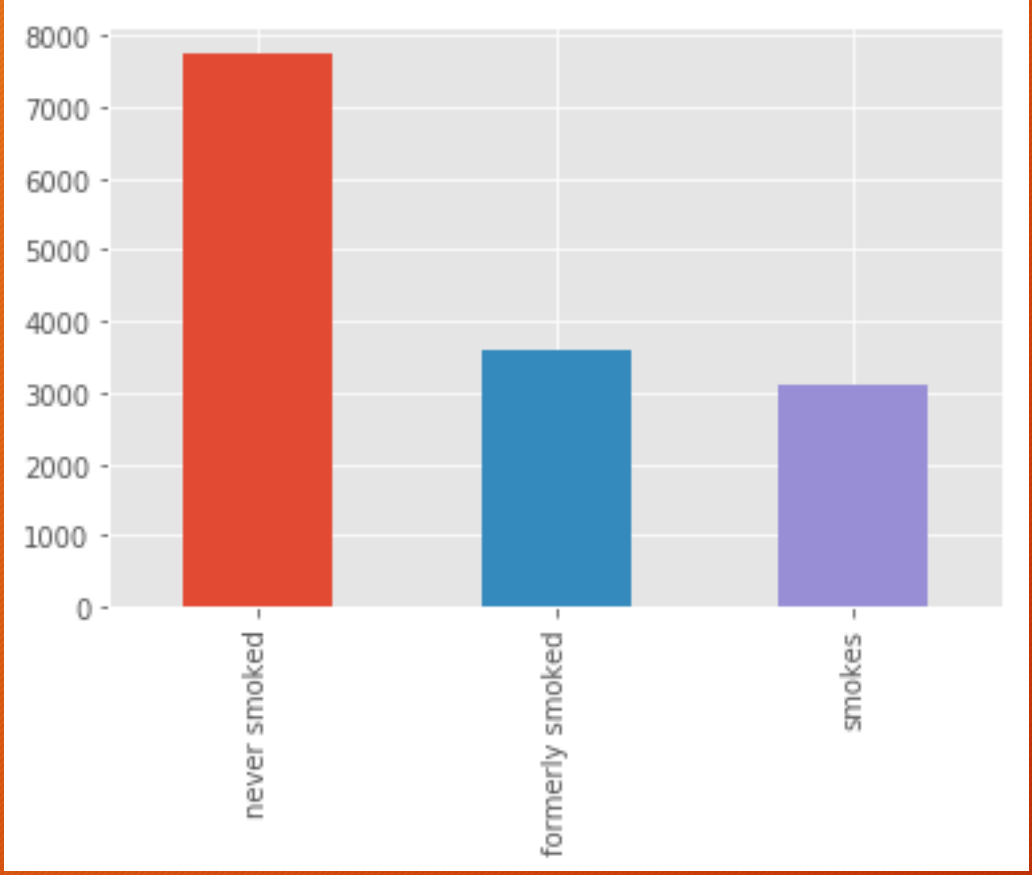
2.9. Подготовка данных. BMI



Outliers пока оставляем без изменений

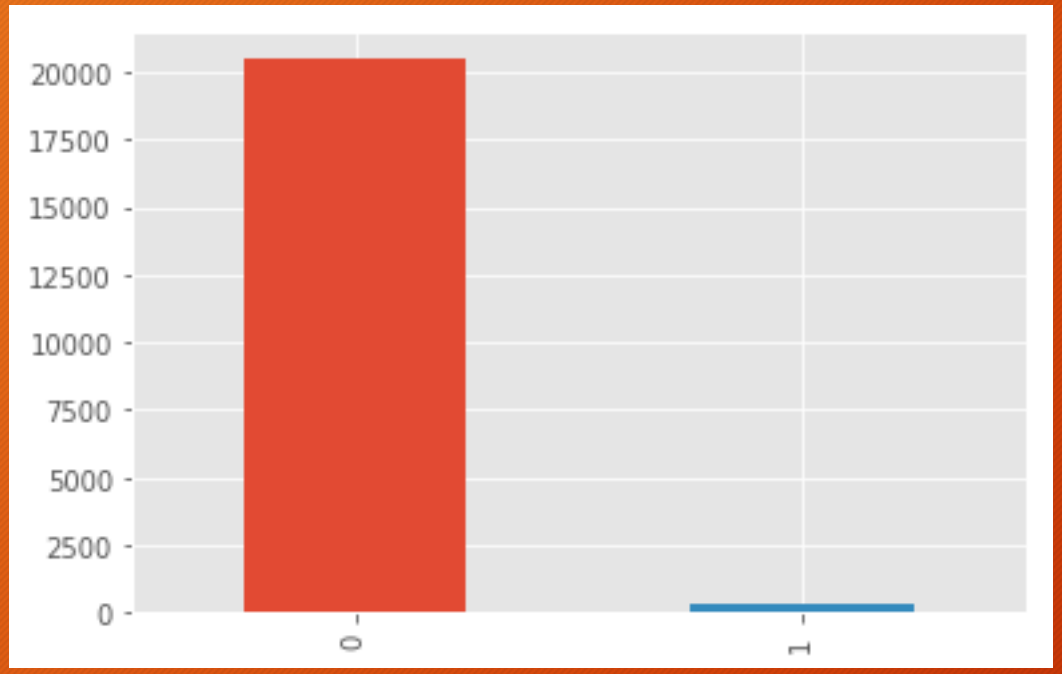
Пропуски заполним средним значением

2.10. Подготовка данных. Smoking status



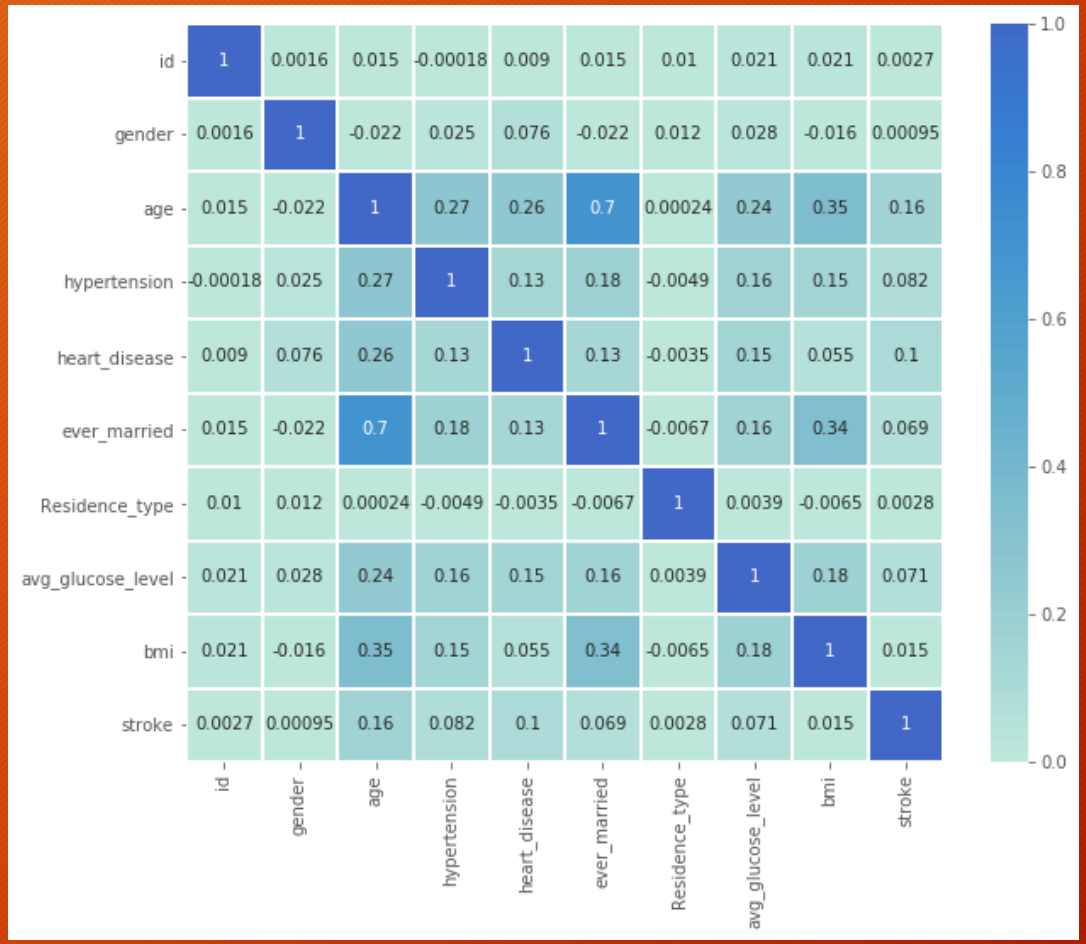
6369 пропусков — заполняем модой

2.10. Подготовка данных. Stroke – целевая функция



Классы сильно несбалансированы

2.11. Подготовка данных. Корреляционная матрица

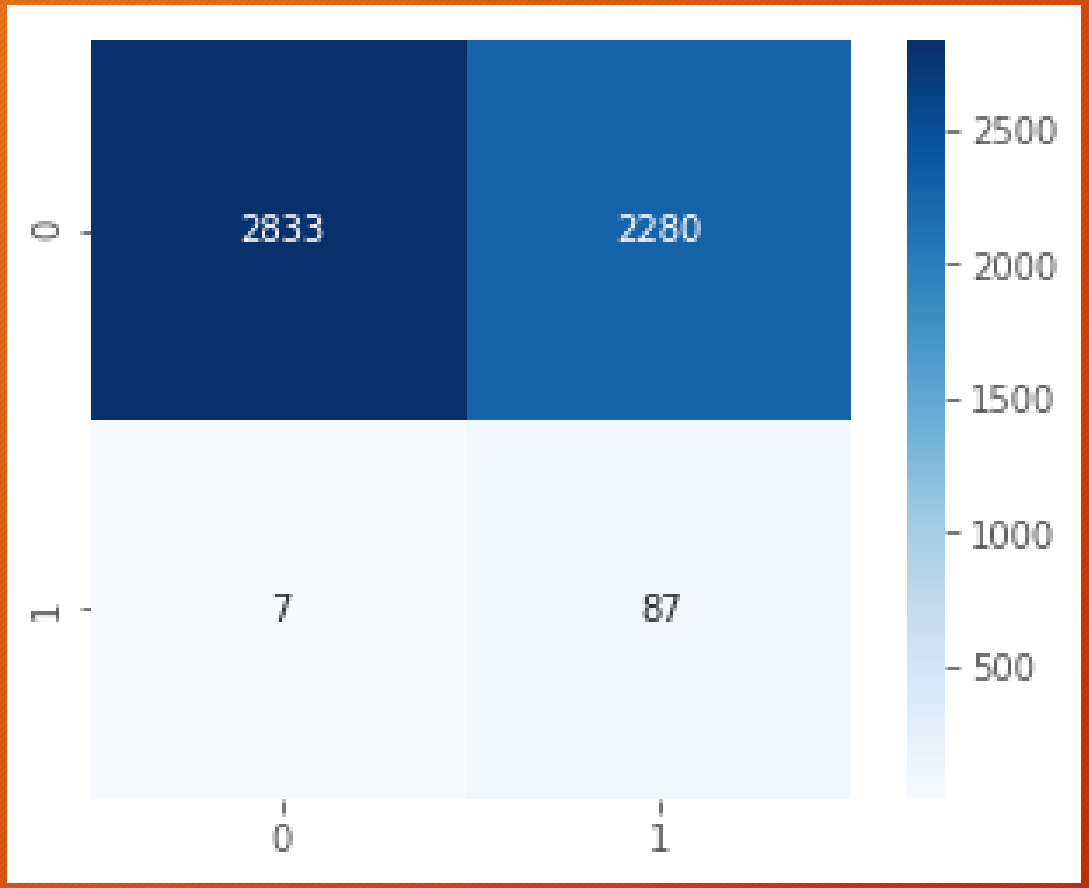


2.12. Подготовка данных. Выбранные признаки

	gender	age	hypertension	heart_disease	Residence_type	avg_glucose_level	bmi
0	0	17.0	0	0	1	133.19	17.8
1	0	8.0	0	0	1	92.53	17.6
2	0	69.0	0	0	1	105.58	26.9
3	0	48.0	0	0	1	67.39	25.6
4	0	10.0	0	0	0	91.60	29.1

- gender
- age
- hypertension
- heart_disease
- Residence_type
- avg_glucose_level
- bmi

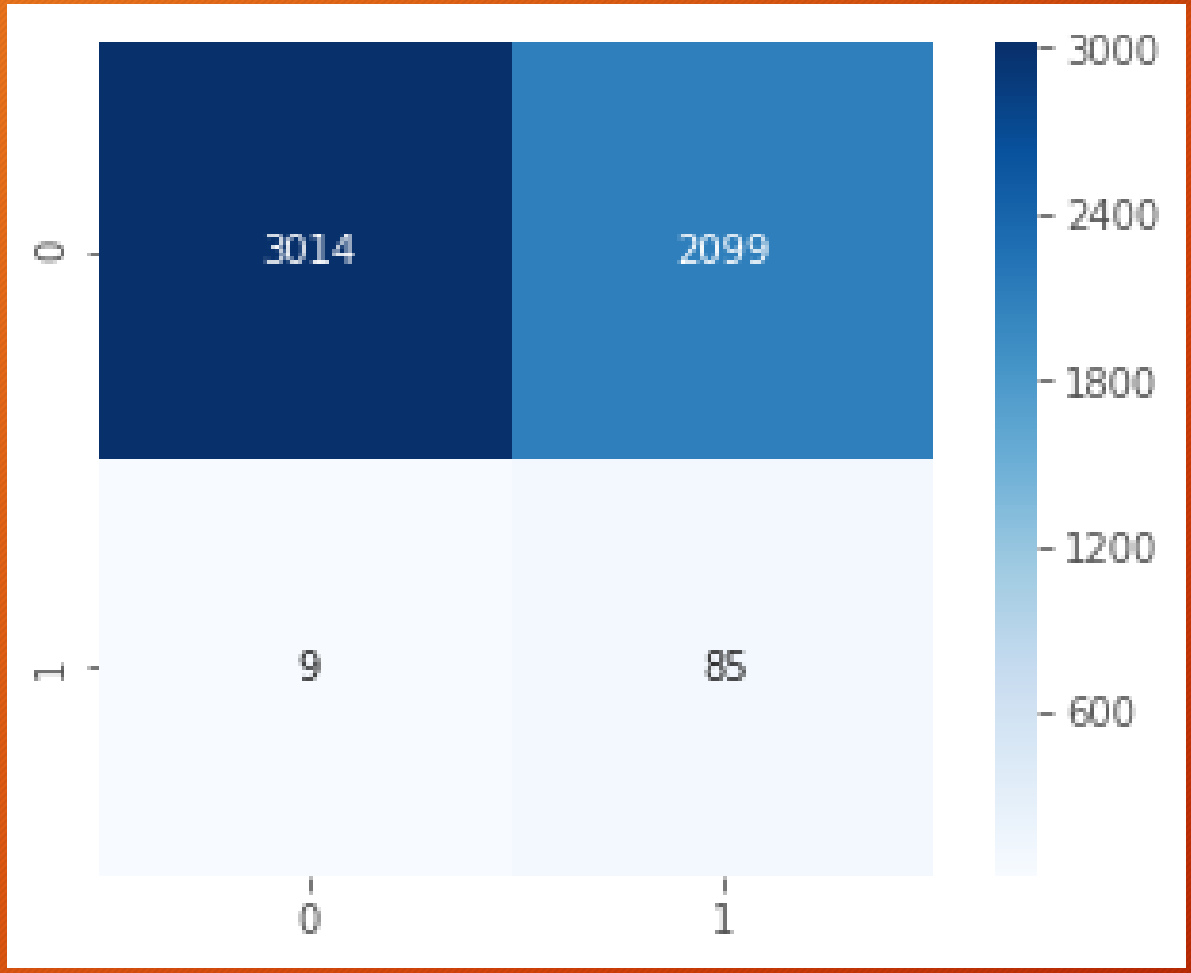
Обучение. Метод опорных векторов



Recall: 0.76596

kernel='linear‘
C=1.0
random_state=228
class_weight='balanced'

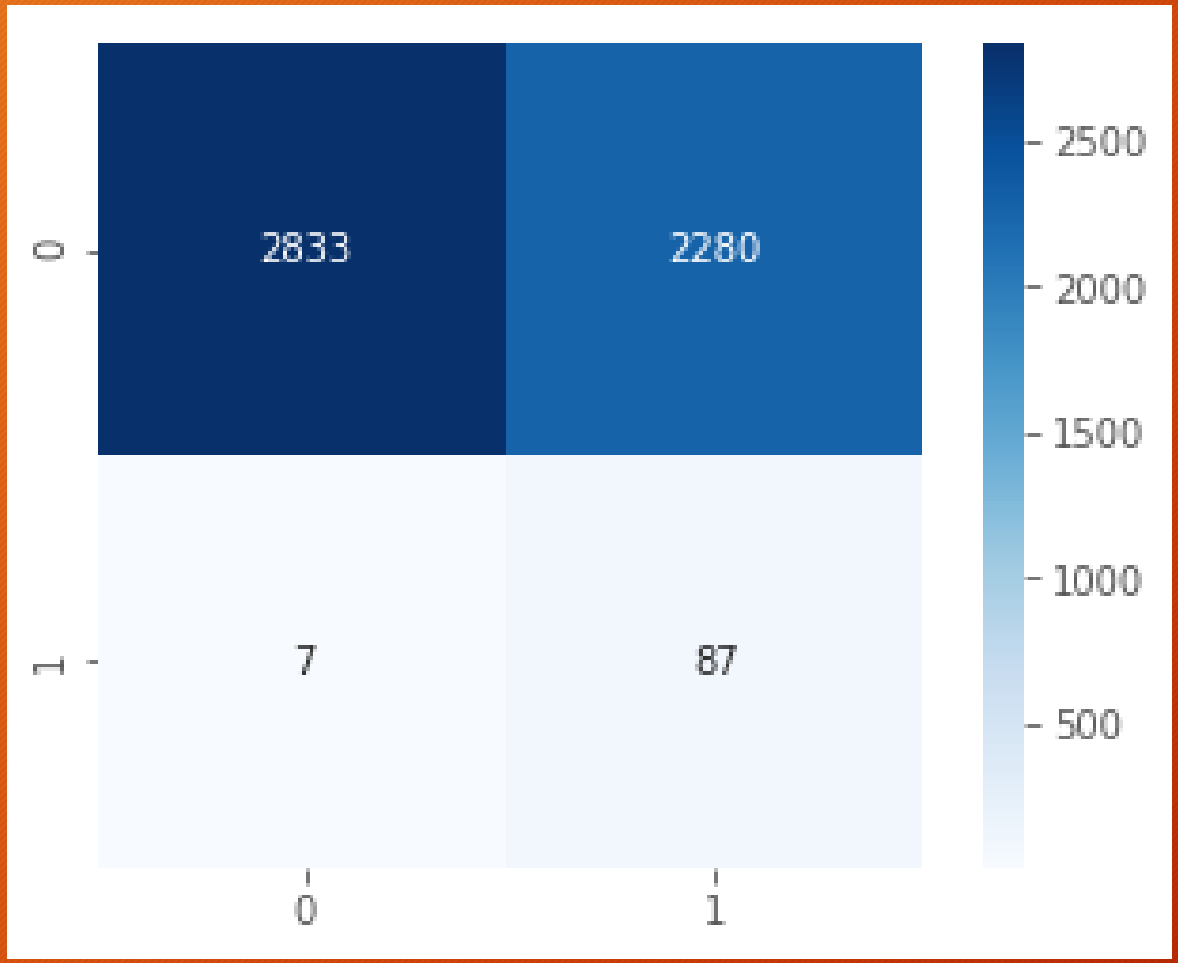
Обучение. Деревья решений



Recall: 0.90426

criterion='entropy‘
max_depth=1
random_state=228
class_weight='balanced'

Обучение. Случайный лес



Recall: 0.92553

n_estimators=2
criterion='entropy‘
max_depth=2
random_state=228
bootstrap=True
class_weight='balanced_subsample'

Обучение со StandardScaler

Random Forest - Recall: 0.90426
Decision Tree - Recall: 0.90426
Adaboost - Recall: 0.00000
SVM - Recall: 0.73404

Обучение с отобранными признаками

	age	hypertension	heart_disease
0	17.0	0	0
1	8.0	0	0
2	69.0	0	0
3	48.0	0	0
4	10.0	0	0

1) age – 0.427157
2) hypertension – 0.243948
3) heart_disease – 0.203331
4) Residence_type – 0.026868
5) avg_glucose_level – 0.025795
6) bmi – 0.019852
7) smoking_status_formerly – 0.019518
8) smoking_status_never smoked – 0.014110
9) smoking_status_smokes – 0.010410

Обучение с отобранными признаками

Random Forest - Recall: 0.90426
Decision Tree - Recall: 0.90426
SVM - Recall: 0.75404

Спасибо за внимание!