Предвидување на срцеви проблеми

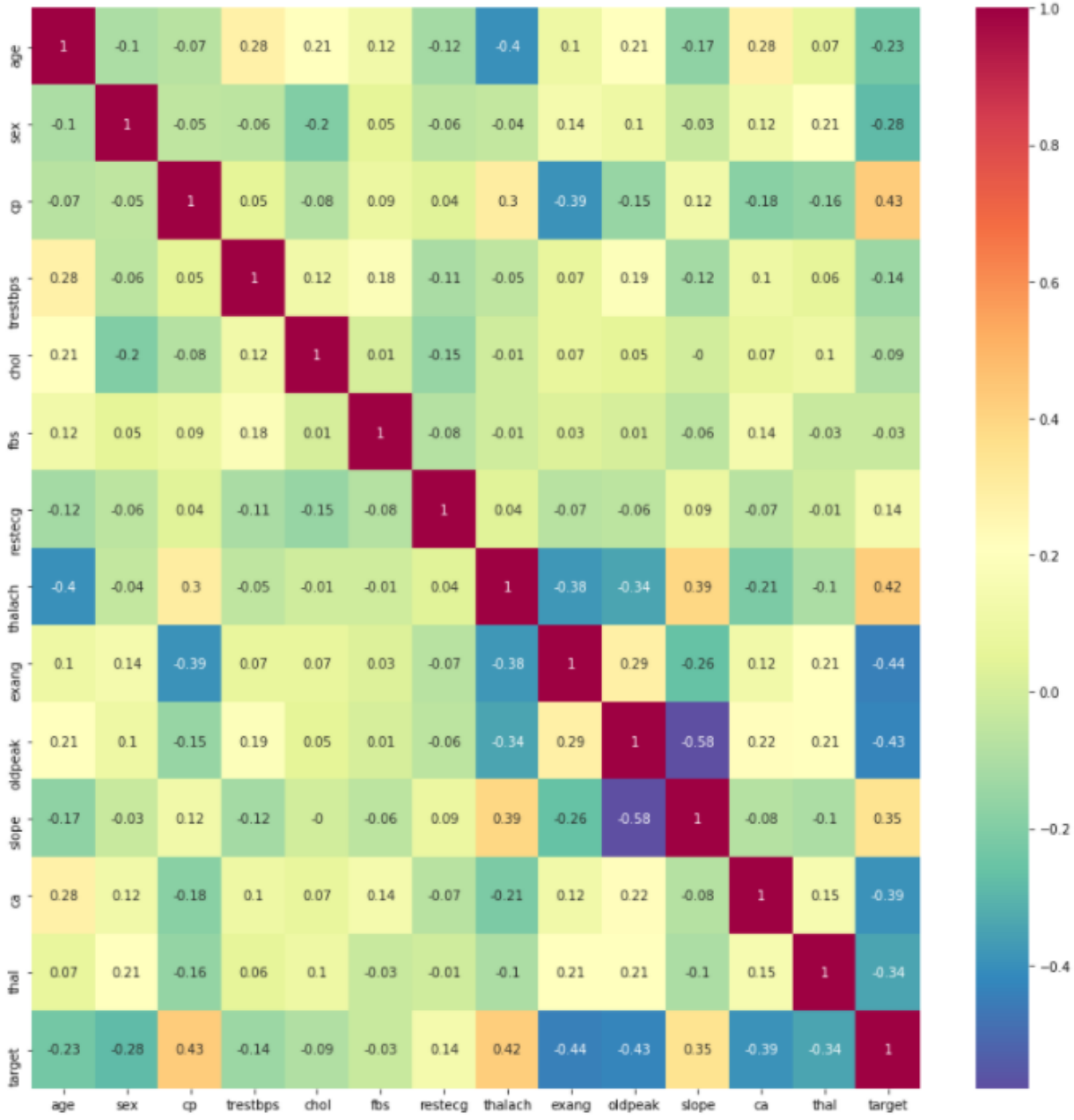
Содржина

Податочно множество

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

- **Chest pain (CP):** Вредности од 0 до 3
- **Trestbps:** крвен притисок во одморена состојба
- **Chol:** холестерол во mg/dl
- **Fbs:** дали има >120 mg/dl шеќер во крв
- **Restecg:** резултати од електрокардиографија
- **Thalach:** максимален пулс
- **Exang**: Болка во гради од вежбање
- **Oldpeak:** ST depression иницирано од вежбање
- **Ca:** број на главни аорти
- **Thal:** дефект на аорти
- **Target: Дали пациентот има срцеви заболувања?**

Зависности помеѓу атрибутите



Визуелизација со PCA

Отстранување на outliers

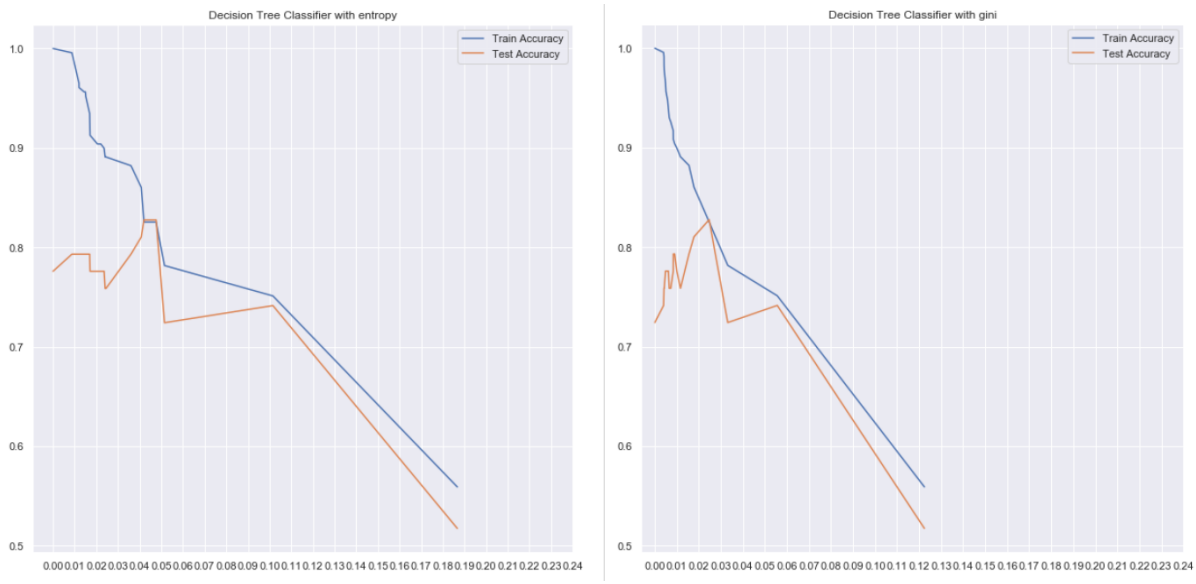
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target	cause of removal
0	65.0	0.0	2.0	140.0	417.0	1.0	0.0	157.0	0.0	0.8	2.0	1.0	2.0	1.0	417.0
1	53.0	0.0	2.0	128.0	216.0	0.0	0.0	115.0	0.0	0.0	2.0	0.0	0.0	1.0	0.0
2	67.0	0.0	2.0	115.0	564.0	0.0	0.0	160.0	0.0	1.6	1.0	0.0	3.0	1.0	564.0
3	52.0	1.0	2.0	138.0	223.0	0.0	1.0	169.0	0.0	0.0	2.0	4.0	2.0	1.0	4.0
4	58.0	1.0	1.0	125.0	220.0	0.0	1.0	144.0	0.0	0.4	1.0	4.0	3.0	1.0	4.0
5	38.0	1.0	2.0	138.0	175.0	0.0	1.0	173.0	0.0	0.0	2.0	4.0	2.0	1.0	4.0
6	38.0	1.0	2.0	138.0	175.0	0.0	1.0	173.0	0.0	0.0	2.0	4.0	2.0	1.0	4.0
7	62.0	0.0	0.0	160.0	164.0	0.0	0.0	145.0	0.0	6.2	0.0	3.0	3.0	0.0	6.2
8	63.0	0.0	0.0	150.0	407.0	0.0	0.0	154.0	0.0	4.0	1.0	3.0	3.0	0.0	407.0
9	55.0	1.0	0.0	140.0	217.0	0.0	1.0	111.0	1.0	5.6	0.0	0.0	3.0	0.0	5.6
10	56.0	0.0	0.0	200.0	288.0	1.0	0.0	133.0	1.0	4.0	0.0	2.0	3.0	0.0	200.0
11	56.0	0.0	0.0	134.0	409.0	0.0	0.0	150.0	1.0	1.9	1.0	2.0	3.0	0.0	409.0
12	54.0	1.0	1.0	192.0	283.0	0.0	0.0	195.0	0.0	0.0	2.0	1.0	3.0	0.0	192.0
13	43.0	1.0	0.0	132.0	247.0	1.0	0.0	143.0	1.0	0.1	1.0	4.0	3.0	0.0	4.0
14	67.0	1.0	0.0	120.0	237.0	0.0	1.0	71.0	0.0	1.0	1.0	0.0	2.0	0.0	71.0
15	52.0	1.0	0.0	128.0	204.0	1.0	1.0	156.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0

⇨ Отстранети се 16 редици од множеството, така што мерката употребена за отфрлање на редици е ако некој атрибут има вредност поголема од 3 * Standard deviation на соодветниот атрибут

Missing values

	Num. of missing values	% of missing values
age	0	0.0
sex	0	0.0
cp	0	0.0
trestbps	0	0.0
chol	0	0.0
fbs	0	0.0
restecg	0	0.0
thalach	0	0.0
exang	0	0.0
oldpeak	0	0.0
slope	0	0.0
ca	0	0.0
thal	0	0.0
target	0	0.0

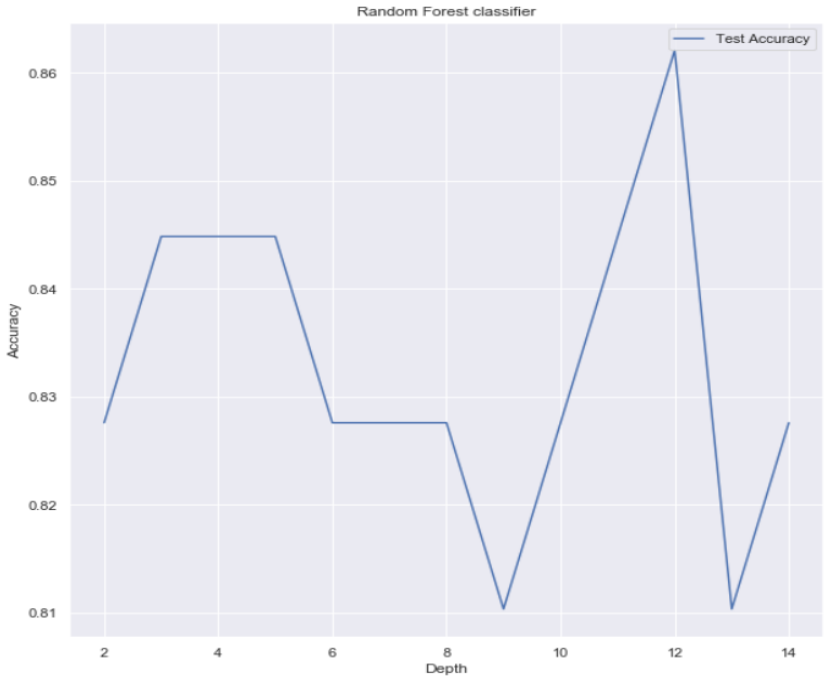
Дрво на одлучување



⇨ Споредба на модели со различна вредност на cpp_alphas и тунирање на истите

```
DTC with entropy precision:  0.9666666666666667
DTC with entropy accuracy:  0.8275862068965517
DTC with gini precision:  0.9666666666666667
DTC with gini accuracy:  0.8275862068965517
```

Random forest classifier



- ⇨ Најдобриот модел е со max depth = 12, со тоа се достигнува вредност на test accuracy со над 0.86

GaussianNB

```
Accuracy score:  0.896551724137931
Precision score:  0.9310344827586207
```

CategoricalNB

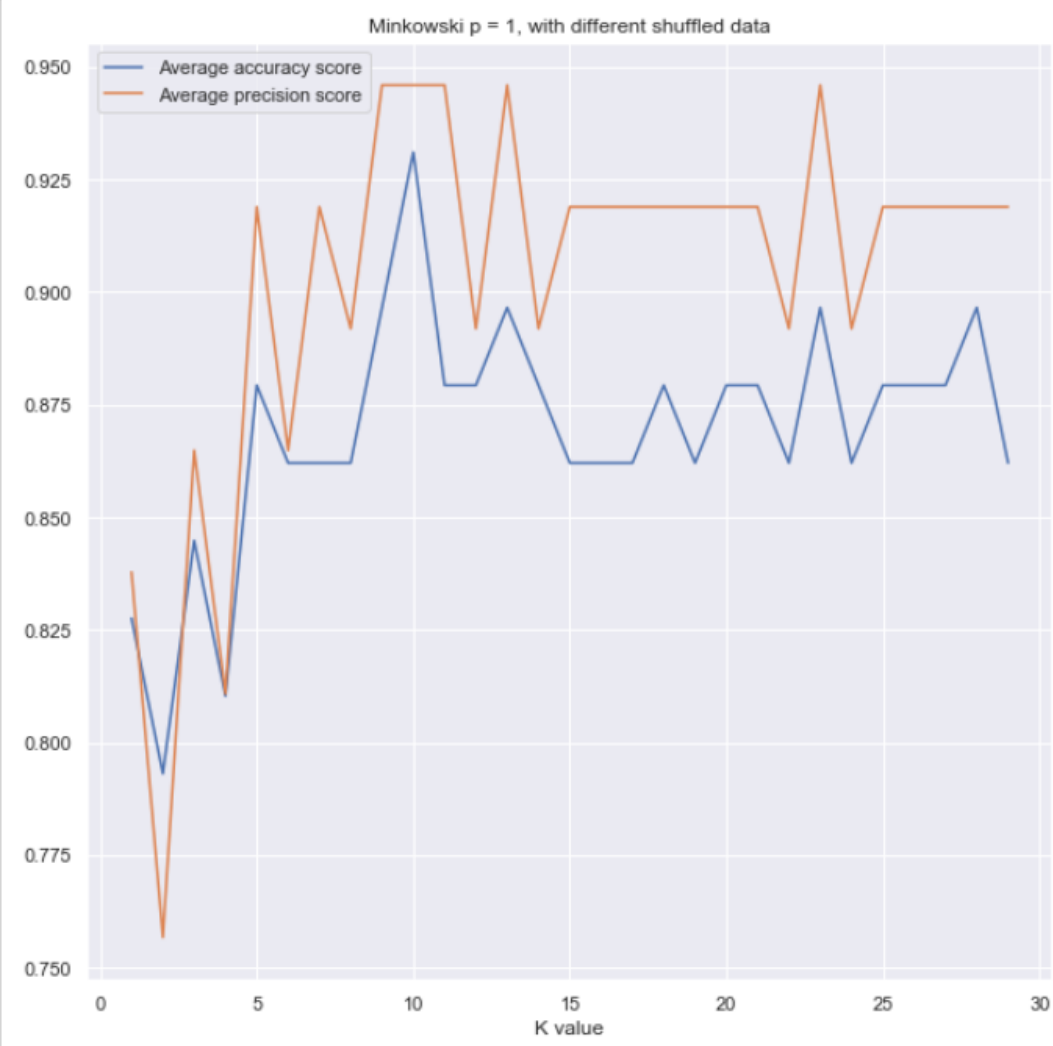
Промена на атрибутите од непрекинат тип според следниве правила и отстранување на атрибутот oldpeak бидејќи има голема зависност со slope атрибутот:

- chol: <150 (normal), >150 <200 (mildly high), >200 <500 (high), >500 (very high)
- age: 18-39 (adult), 40-59 (middle aged), >60 (senior adult)
- trestbps (Resting blood pressure, Systolic): <120 (normal), >120 <140 (elevated), >140 <160 (High blood pressure), >160 (Hypertension)
- thalach (Maximum hearth rate): <120 normal, >120 <150 moderate, >150 high
- oldpeak: removed

Се добиваат следниве резултати каде што има мало подобрување во accuracy:

```
Accuracy score:  0.9137931034482759
Precision score:  0.896551724137931
```

K-Neighbors Classifier



- ⇨ Тестирање на различни модели, така што за секоја K вредност употребени се 100 различни варијанти на множество за да биде појасно кој е најдобрата вредност на K

Со одбирање на к=10 и користење на Minkowski мерка за дистанца (p=1) се добиваат следниве резултати

	precision	recall	f1-score	support
0	0.90	0.90	0.90	21
1	0.95	0.95	0.95	37
accuracy			0.93	58
macro avg	0.93	0.93	0.93	58
weighted avg	0.93	0.93	0.93	58

```
[[19  2]
 [ 2 35]]
```

Предвидување на холестерол во крвта со помош на линеарна регресија

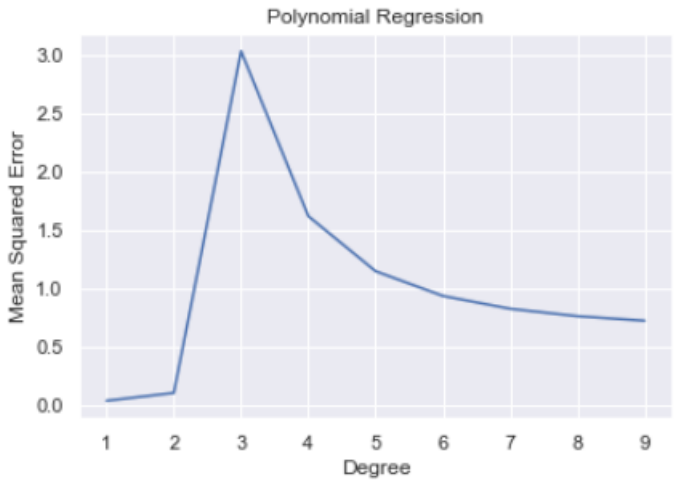
Резултати на обична линеарна регресија:

```
Mean Squared Error:  2322.0206946198655
Mean absolute Error:  38.153844251238475
Root Mean squared error:  48.18734994394136
```

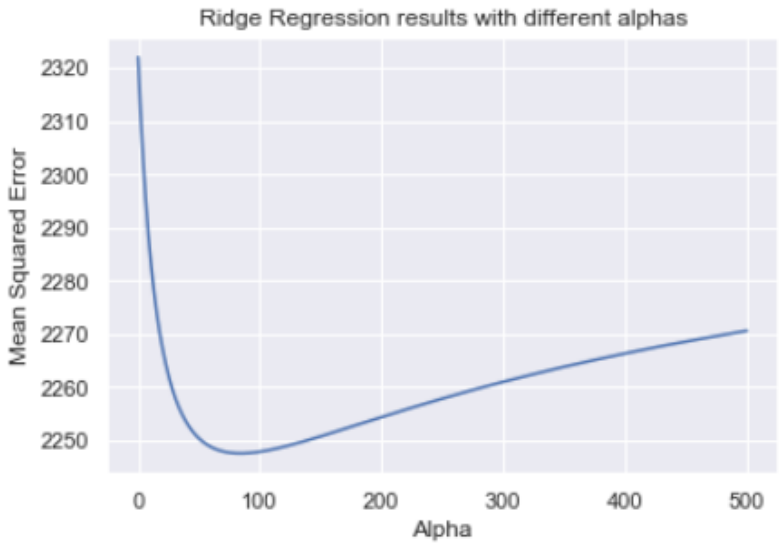
Резултати на обична линеарна регресија со скалирање на променливите:

```
Mean Squared Error:  0.032329314638837536
Mean absolute Error:  0.14236509048969584
Root Mean squared error:  0.17980354456694544
```

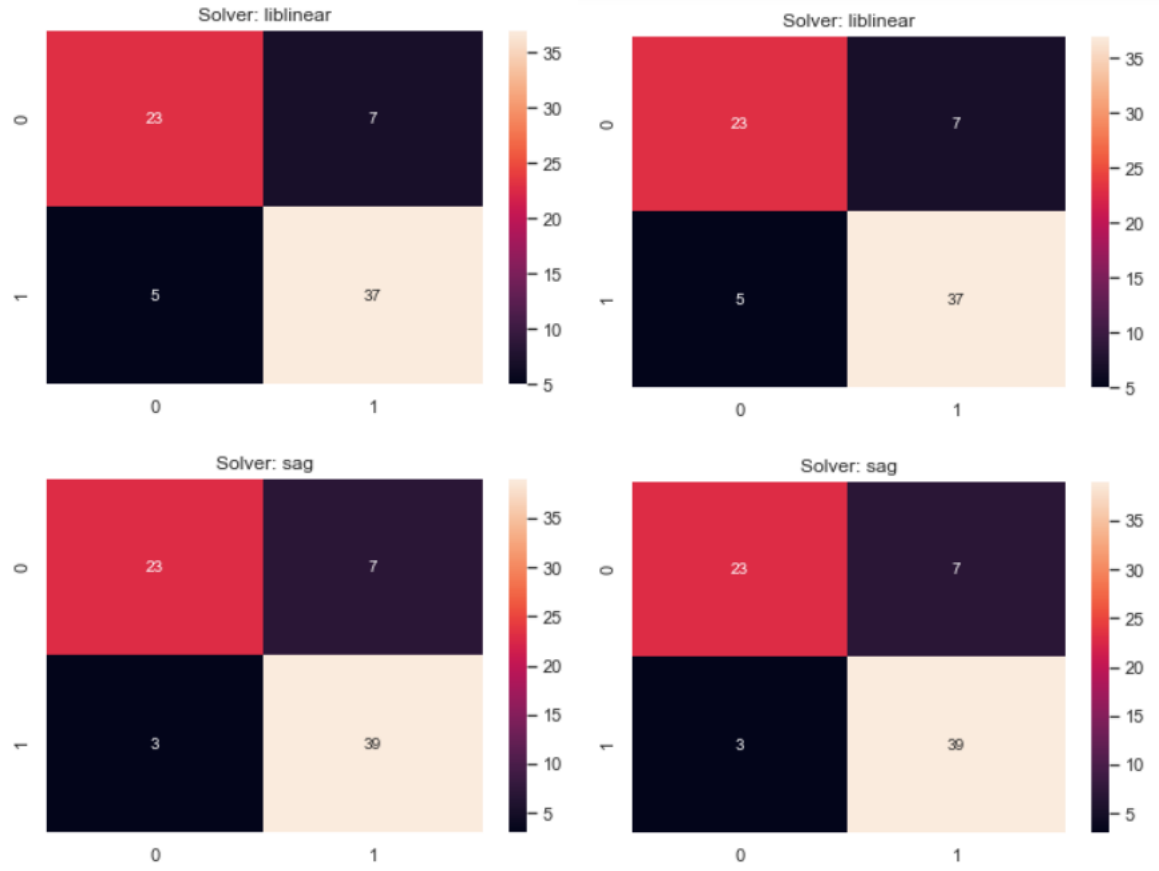
Полиномна регресија со скалирање:



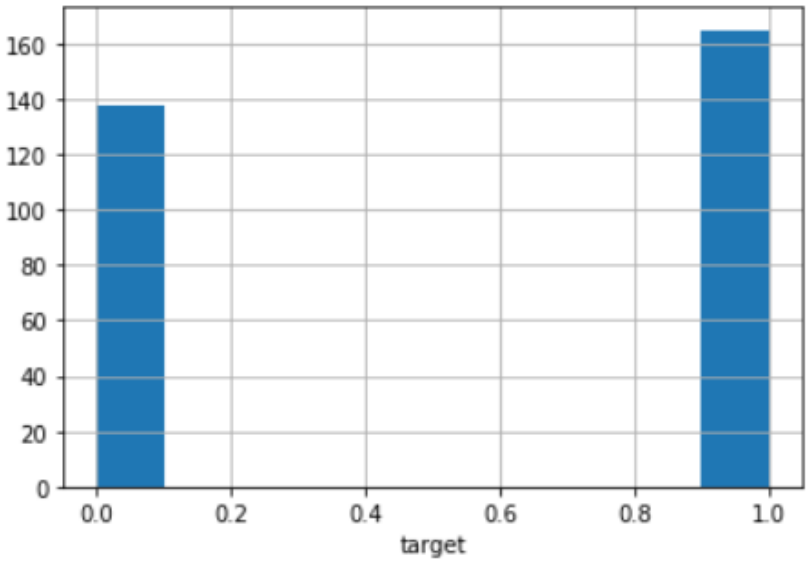
Ridge регресија



Предвидување на target атрибутот со помош на логистичка регресија



LDA со не балансирано и балансирано множество



	precision	recall	f1-score	support
0	0.82	0.85	0.84	27
1	0.91	0.89	0.90	45
accuracy			0.88	72
macro avg	0.87	0.87	0.87	72
weighted avg	0.88	0.88	0.88	72

Не балансирано

	precision	recall	f1-score	support
0	0.87	0.89	0.88	38
1	0.90	0.88	0.89	41
accuracy			0.89	79
macro avg	0.89	0.89	0.89	79
weighted avg	0.89	0.89	0.89	79

Балансирано

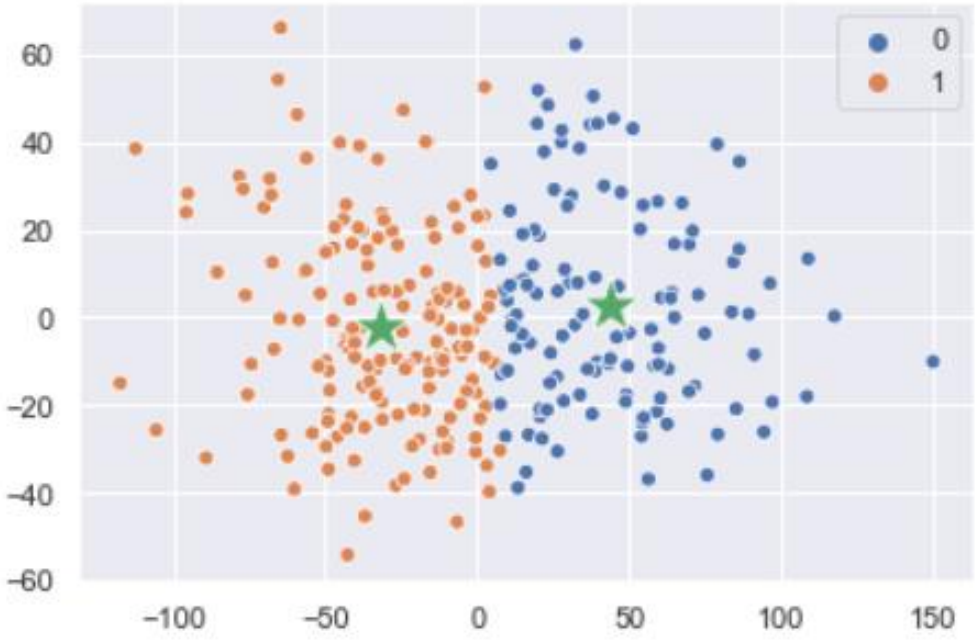
Ensemble методи

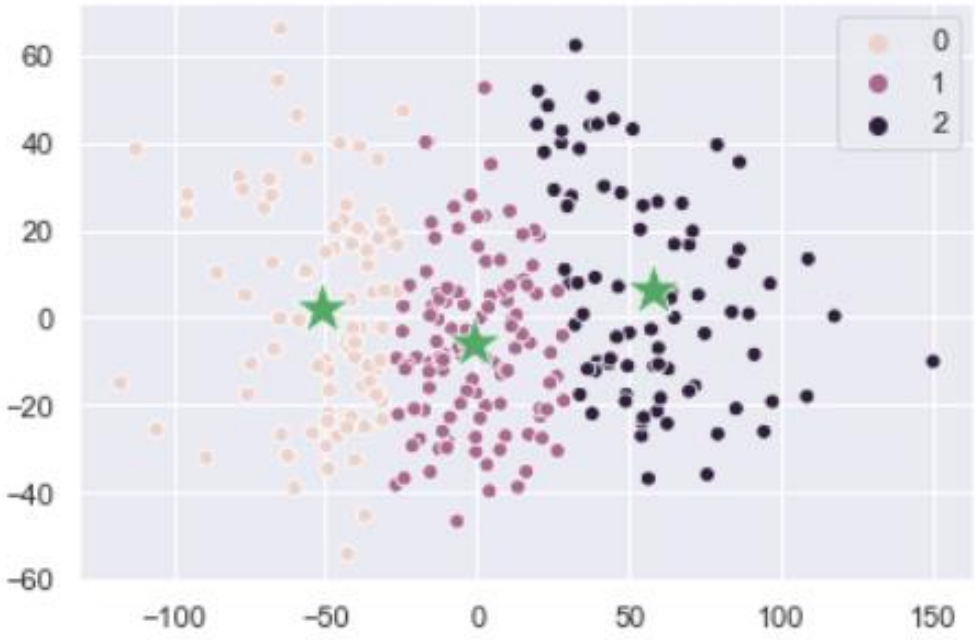
Со помош на 3 различни модели (Логистички регресионен модел, Random forest classifier и KNeighborsClassifier) се добива Test Accuracy:

```
0.9137931034482759
```

Кластер модели

KMeans





Algomerative кластерирање