



Решавање проблема из области класификације

- ◎ Разврставање електронске поште
- ◎ *Стабло одлучивања*
 - интерпретабилност
 - идентификација најзначајнијег атрибута као и везе између једног или више атрибута
 - применљив и на регресију и на класификацију
 - преприлагођавање

Да ли може боље?

1

Ансамбл методе

Решава се проблем преприлагођавања



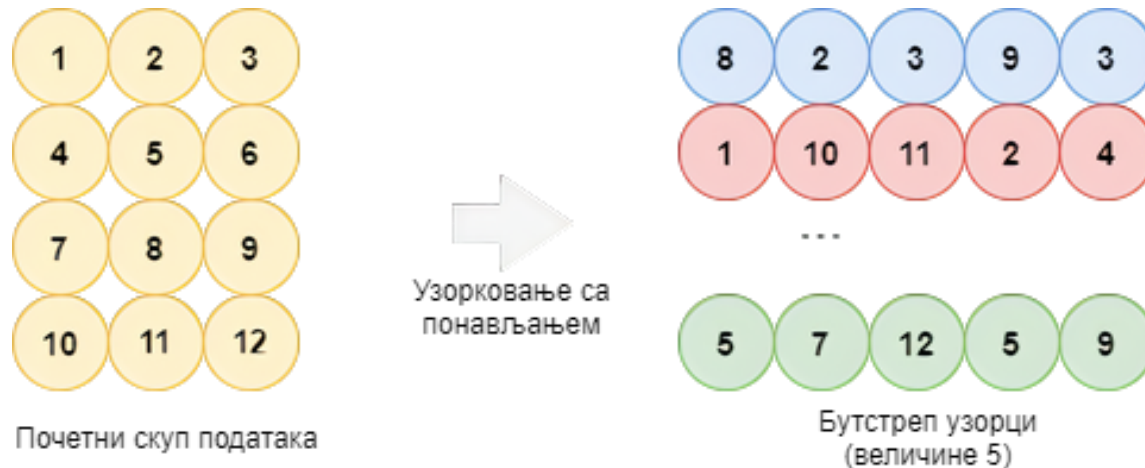
Ансамбл методе

- Из скупа података за обучавање се **случајно** одаберу **скупови података за обучавање сваког појединачног модела**. На тај начин се генерише више различитих модела, а затим се комбинацијом предикција појединачних модела формира предикција ансамбла.

2 Паковање



Бутстрепинг



1.

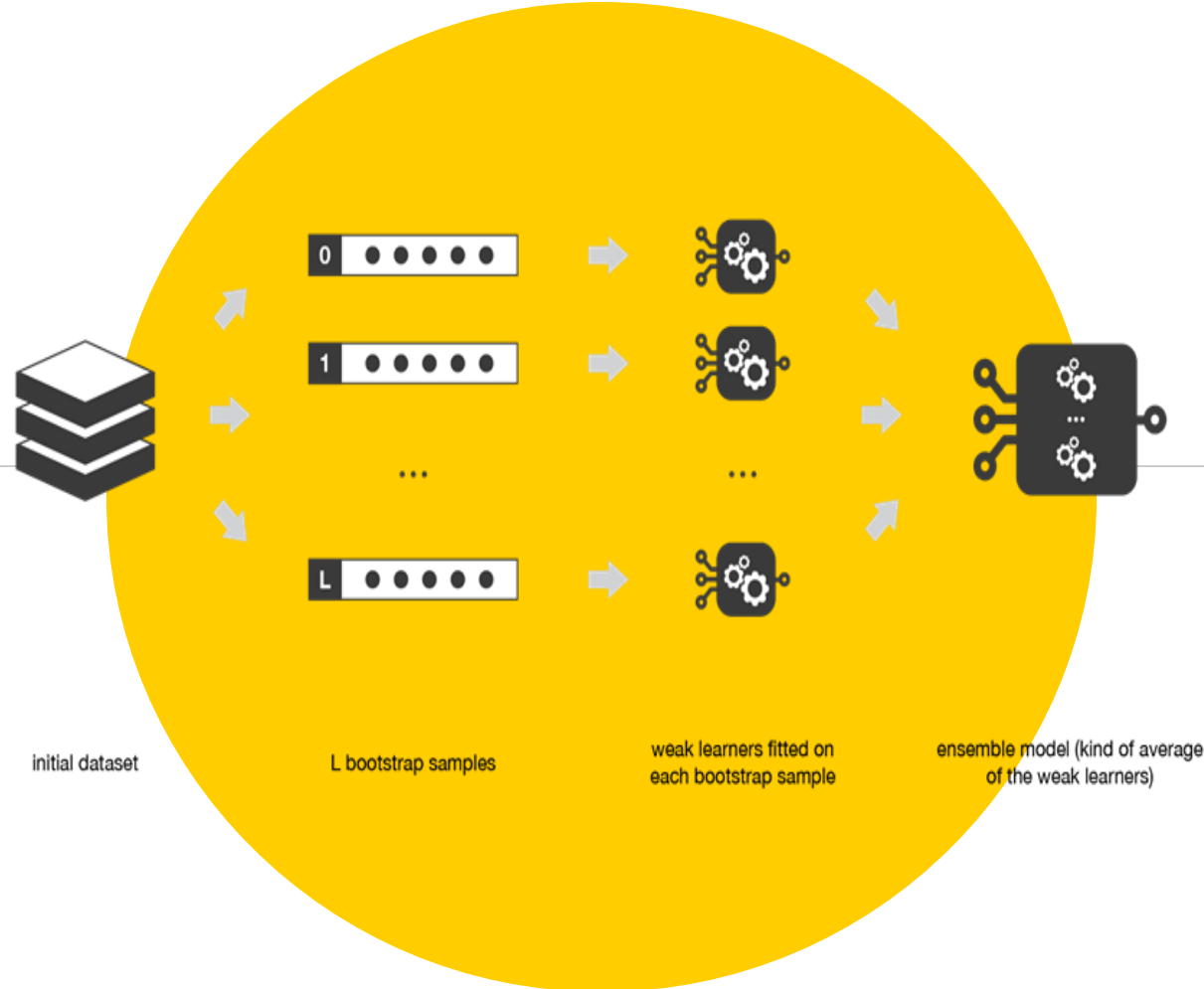
скоро репрезентативни
и скоро независни
узорци (енгл. *almost i. i.
d. samples, independent
identically distributed*)

2.

$\rho\sigma^2 + \frac{1-\rho}{M}\sigma^2$ –
просечна варијанса M
таквих узорака, сваки
са варијансом σ^2

3.

мање од $\frac{2}{3}$ примерака
се појављује у једном
бутстреп узорку



Учење ансамбла паковањем

- Гласање ансамбла
- Упросечавање предикција



Гласање ансамбла

Чврсто гласање (енгл. **Hard voting**)

$$\psi_{\mathcal{L}, \theta_1, \dots, \theta_M}(x) = \operatorname{argmax}_{c \in \mathcal{Y}} \sum_{m=1}^M \mathbf{1}_{(\varphi_{\mathcal{L}, \theta_m}(x)=c)}$$

Класификатор	Предикција
Класификатор 1	Класа 1
Класификатор 2	Класа 1
Класификатор 3	Класа 0

- *класа 1* је предвиђање ансамбла

Меко гласање (енгл. **Soft voting**)

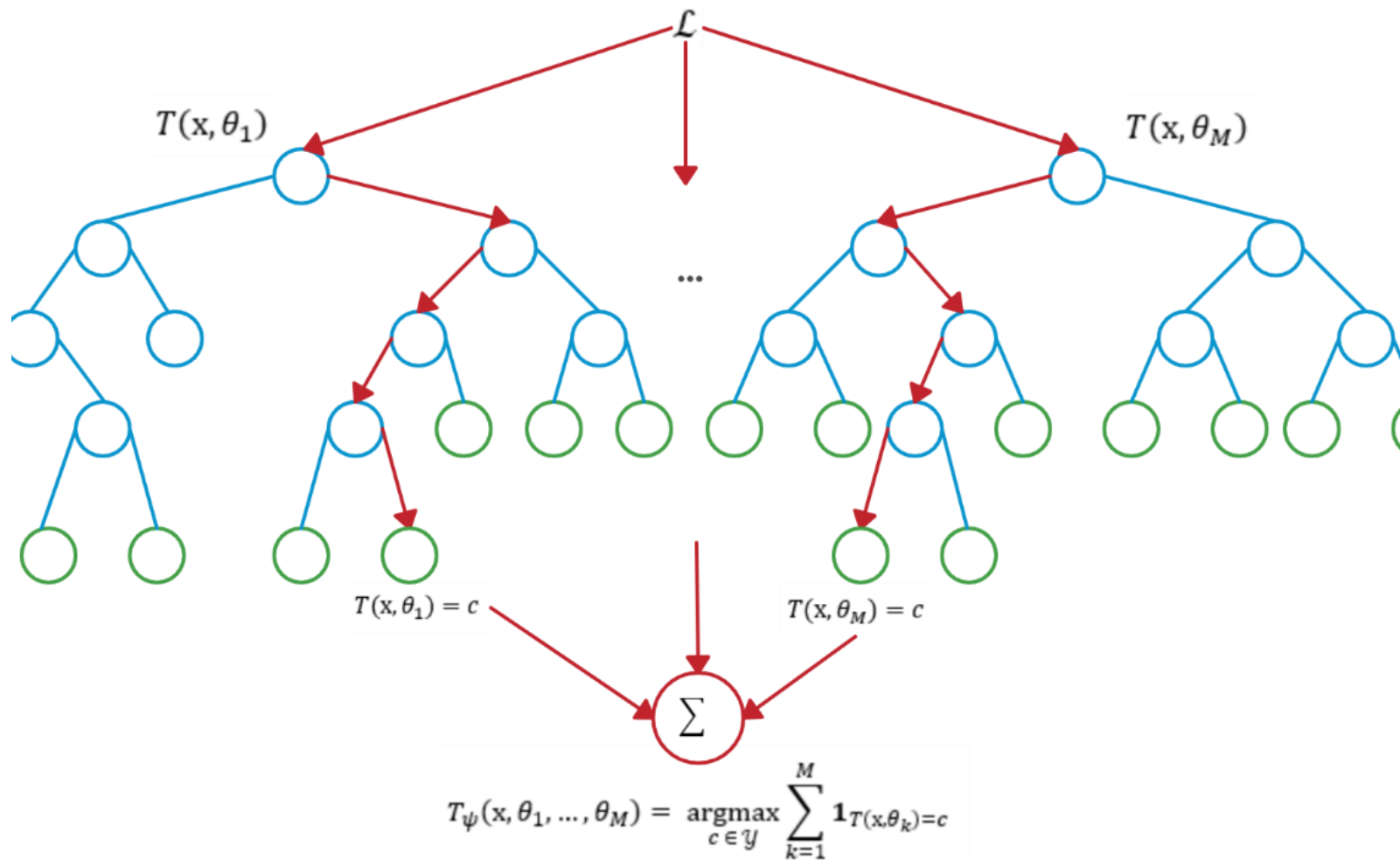
$$\psi_{\mathcal{L}, \theta_1, \dots, \theta_M}(x) = \operatorname{argmax}_{c \in \mathcal{Y}} \frac{1}{M} \sum_{m=1}^M \hat{p}_{\mathcal{L}, \theta_m}(Y = c \mid X = x)$$

Класификатор	Класа 0	Класа 1
Класификатор 1	0,3	0,7
Класификатор 2	0,5	0,5
Класификатор 3	0,4	0,6

- За *класу 0* просечна вероватноћа: $\frac{0,3 + 0,5 + 0,4}{3} = 0,4$
- За *класу 1*: $\frac{0,7 + 0,5 + 0,6}{3} = 0,6$
- *класа 1* је предвиђање ансамбла

3

Насумична шума (енгл. *Random Forest*)



Насумична
класифікаціона
шума

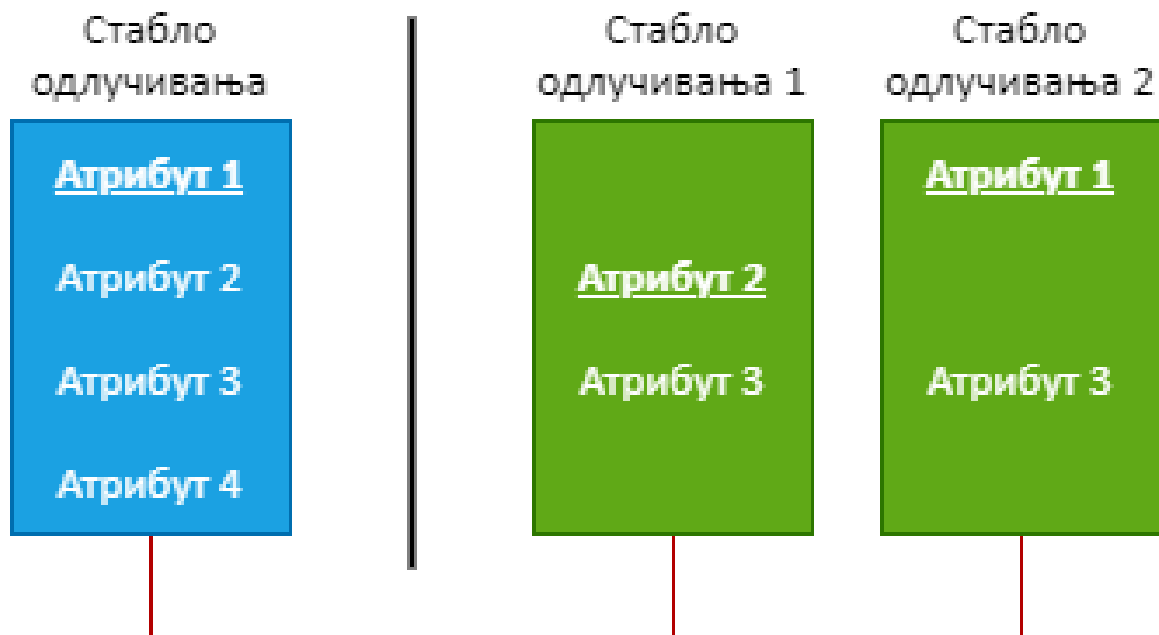
Насумична шума је класификатор који се састоји из колекције стабала одлучивања $\{T(x, \theta_k), k = 1, \dots, M\}$ где су $\{\theta_k\}$ независни идентично дистрибуирани насумични вектори, а свако стабло даје један глас за најпопуларнију класу којој припада инстанца x .



“



Избор атрибута који ће се наћи у чворовима одлуке



Ако атрибута има p , бира се неко $t \ll p$, тако да се за сваки чвор насумично бира t атрибута од p и у чвору ће се наћи најзначајнији атрибут од тих t .

На грешку предвиђања утичу две ствари:

1. Веза између било која два стабла у шуми
2. Јачина сваког индивидуалног стабла

Оптимална вредност t ?

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Hot	High	Weak	Yes
Windy	Cold	Low	Weak	Yes



Out of Bag скуп

Око $1/3$ инстанци се не појављује у бутстреп узорку.

Изостављена врста за стабла 1, 3 и 5.

Outlook	Temperature	Humidity	Wind	Play Tennis
Windy	Cold	Low	Weak	Yes

Стабло одлучивања (k)	Предикција
1	YES
3	NO
5	YES
Већински глас: YES	

- OOB скор (енгл. **OOB score**)
- OOB грешка предвиђања ε_{oob} (енгл. **OOB error**)

$$(x_i, y_i) \notin \mathcal{L}_k$$

$$T_\psi(x_i, \theta_k) = \hat{y}_i$$

$$\varepsilon_{oob} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{(y_i \neq \hat{y}_i)}$$



Важност атрибута

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Survived
1	3	male	22	1	0	7.25	S	0
2	1	female	58	0	0	26.55	S	1
3	2	male	35	0	0	26	S	0
...	...	...	...	...	...	...	...	...
100	3	male	7	4	1	7.8	S	0
101	2	female	34	0	1	23	S	1
...	...	...	...	...	...	...	...	...
891	1	male	26	0	0	30	C	1

Средње умањење прецизности (енгл. *Mean Decrease in Accuracy – MDA*)

Израчунати
OOB грешку.

Пермутовати
вредности у
само једној
колони у OOB
скупу.

Израчунати
на том
скупу OOB
грешку и
упоредити
их.

Поништити
пермутовање
из 2. корака.
Поновити
корак 2 на
другом
атрибуту.

```

[1]  function RandomForest( $L, F$ )
[2]       $T_\psi \leftarrow \emptyset$ 
[3]      for  $i \in 1, \dots, M$  do
[4]           $L^{(i)} \leftarrow$  бутстреп узорак скупа  $L$ 
[5]           $T_i \leftarrow$  RandomizedTreeLearn( $L^{(i)}, F$ )
[6]           $T_\psi \leftarrow T_\psi \cup \{T_i\}$ 
[7]      end for
[8]      return  $T_\psi$ 
[9]  end function
[10]
[11] function RandomizedTreeLearn( $L, F$ )
[12]     За сваки чвор:
[13]          $f \leftarrow$  јако мали подскуп скупа  $F$ 
[14]         Искористи у чвору најзначајнији атрибут из  $f$ 
[15]         и разврстај примере на основу његове вредности
[16]     return обучено стабло
[17] end function

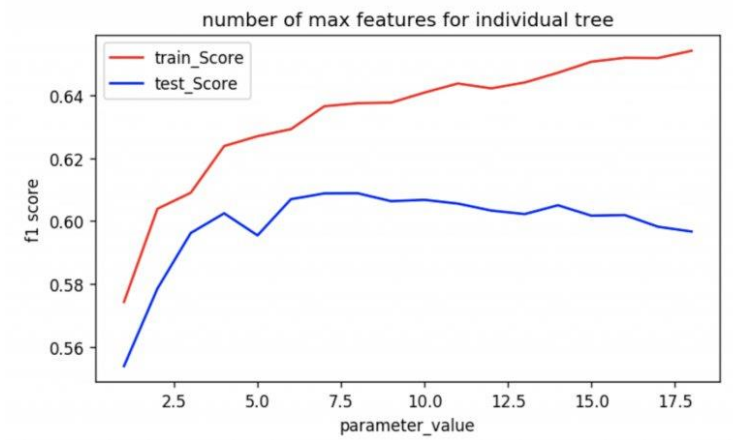
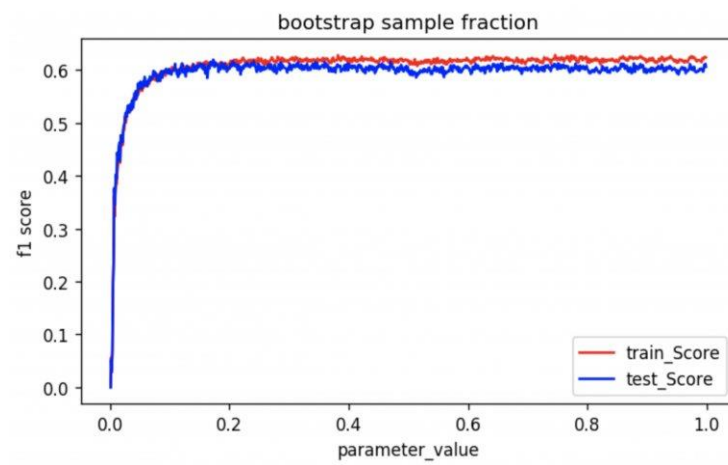
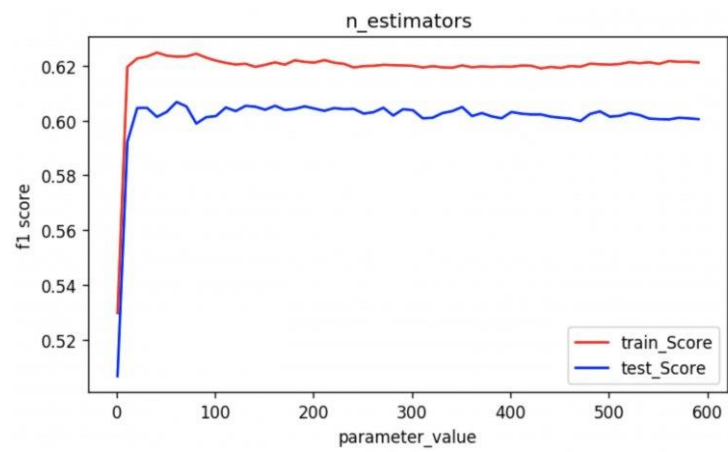
```

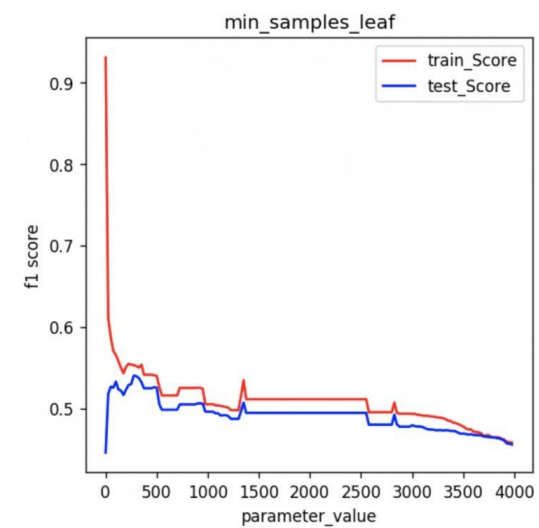
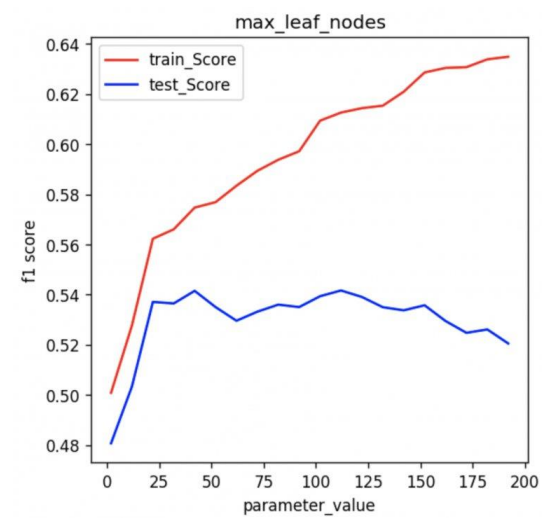
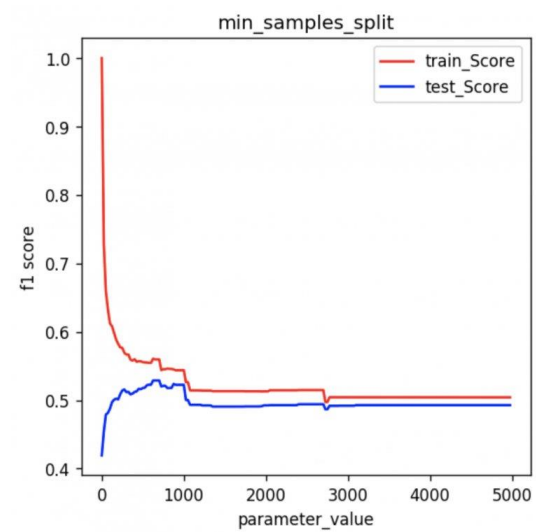
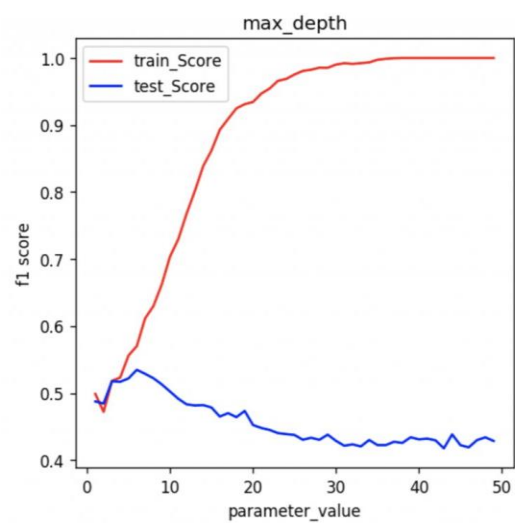
Алгоритам насумичне шуме

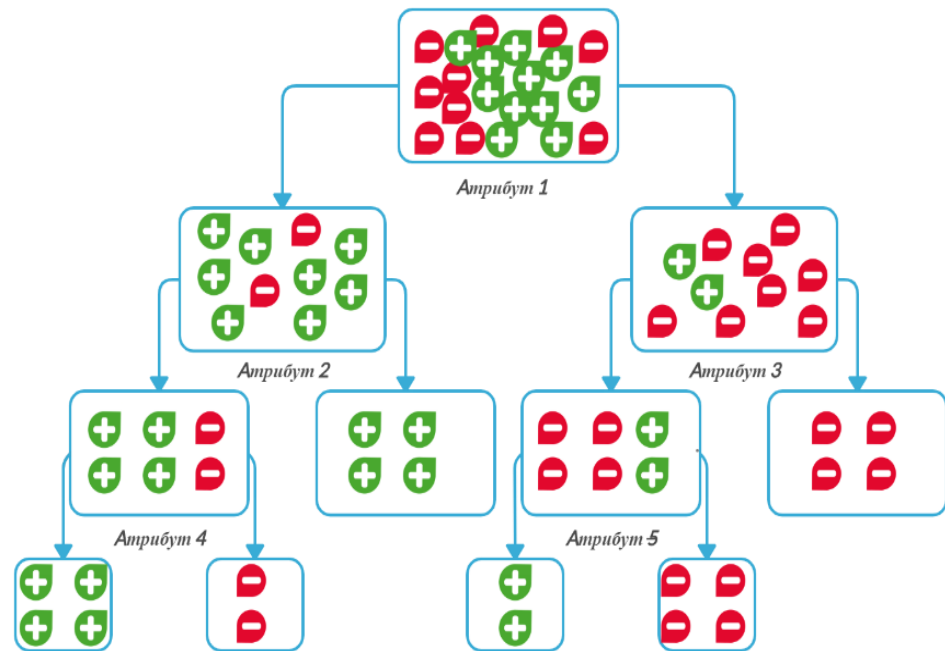


4

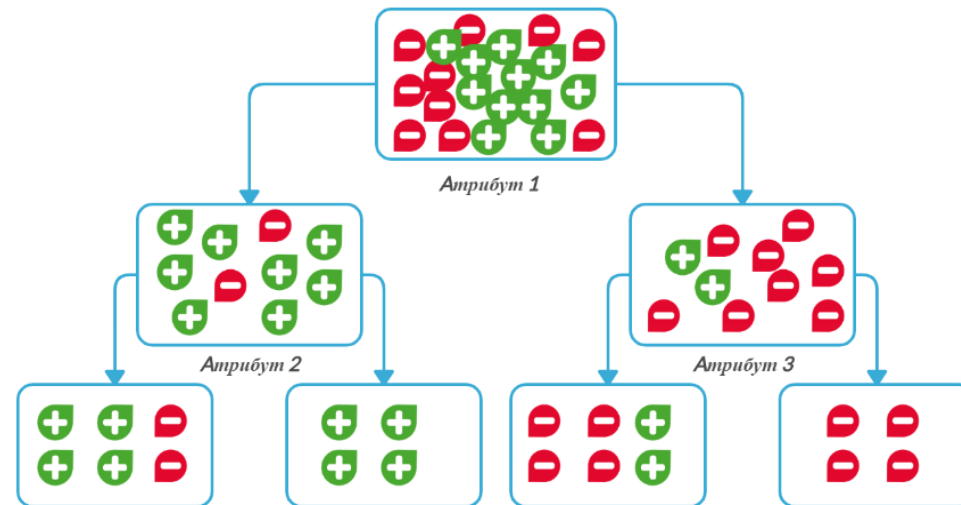
Хиперпараметри на сумичне шуме



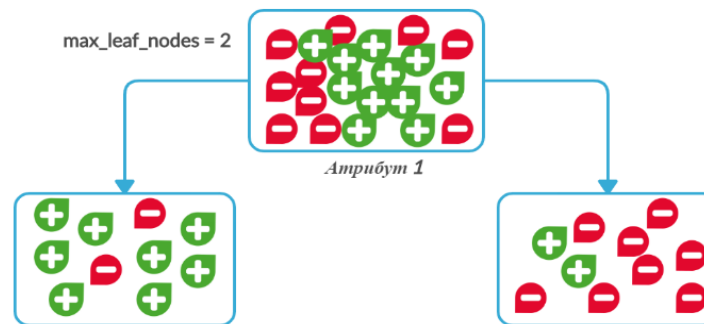




$min_samples_split = 2$

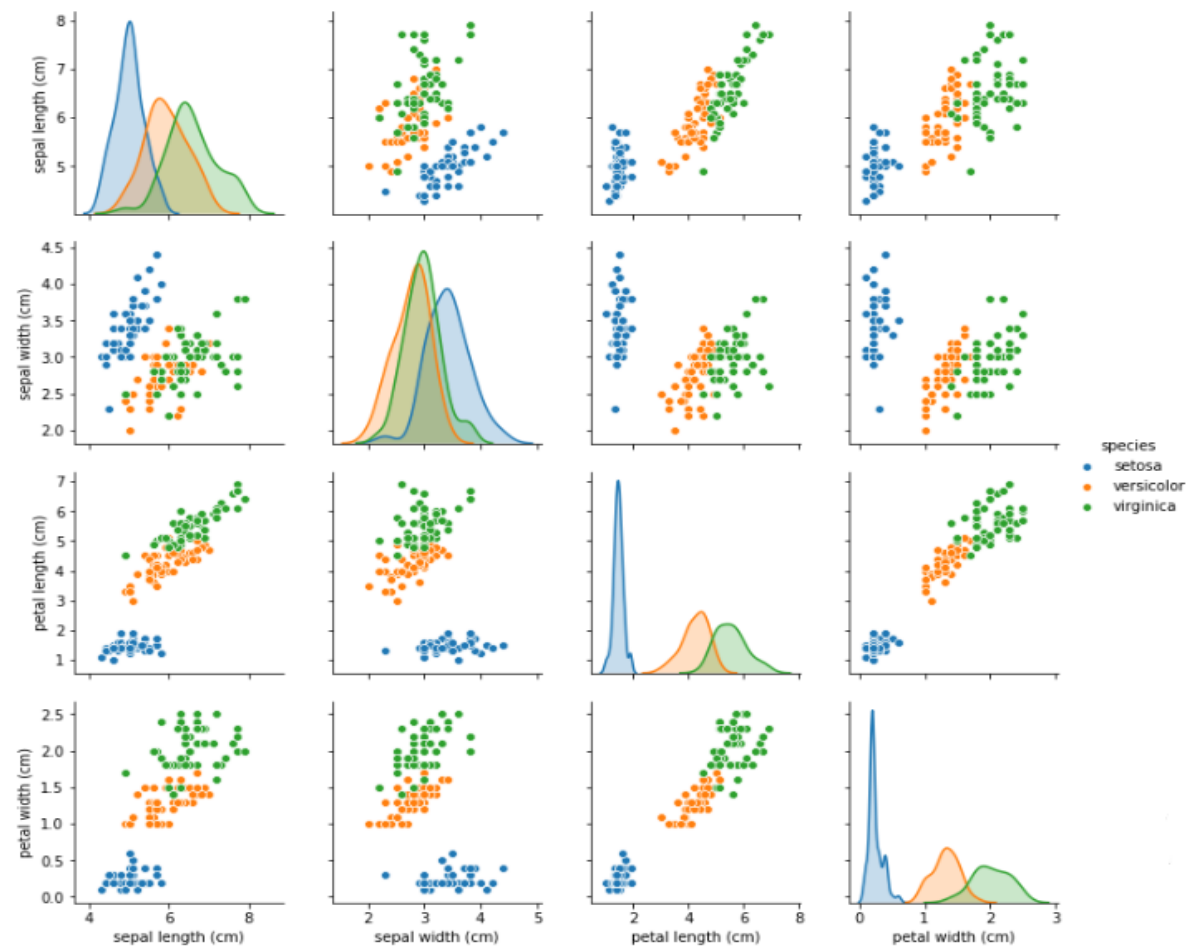


$min_samples_split = 6$



$max_leaf_nodes = 2$





**Креирање модела насумичне шуме за
тренинг скуп *Iris***



```
from sklearn.model_selection import train_test_split
```

```
# upotreba stratify obezbedjuje da oba skupa sadrze reprezentativne instance sve
```

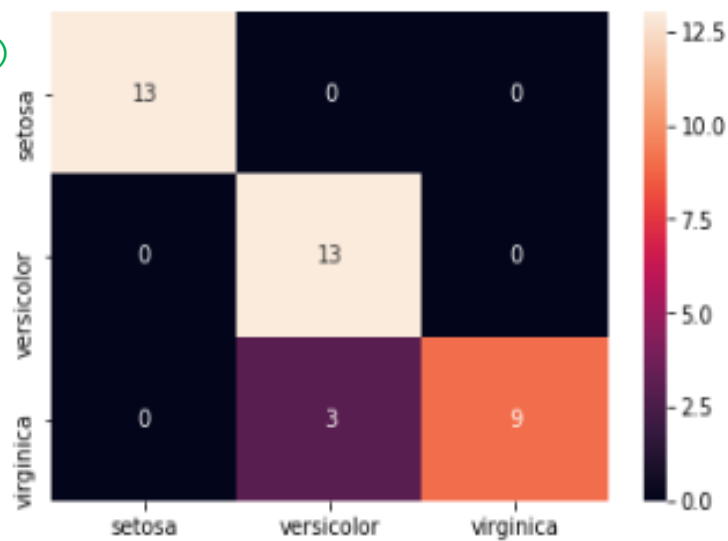
```
# tri klase
```

```
X_train, X_test, y_train, y_test = train_test_split(df[iris.feature_names],  
iris.target, test_size=0.25, stratify=iris.target, random_state=123456)
```

```
from sklearn.ensemble import RandomForestClassifier
```

```
rf = RandomForestClassifier(n_estimators=100, oob_score=True, random_state=123456)
```

```
rf.fit(X_train, y_train)
```



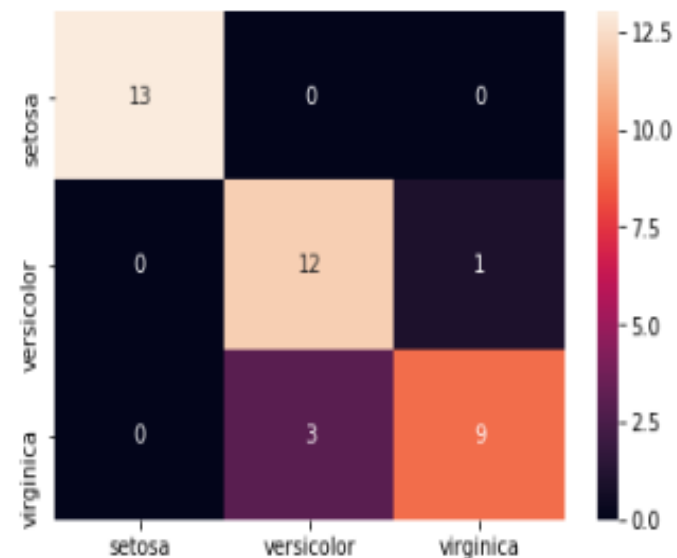
Out-of-bag score estimate: 0.946

Mean accuracy score: 0.921

```
from sklearn.tree import DecisionTreeClassifier
```

```
clf = DecisionTreeClassifier()
```

```
clf.fit(X_train, y_train)
```



Mean accuracy score: 0.895

Насумична шума vs. Стабло одлучивања



Регресија алгоритмом насумичне шуме

