

# Bayesova klasifikacija

---

Luka Jerman, 10.8.2022

# Vsebina

1. Predstavitev klasifikatorja .....	3
2. Predstavitev podatkovne zbirke .....	3
3. Ovrednoteni rezultati klasifikacijskih algoritmov .....	4
3.1. Točnost.....	4
3.2. Priklic.....	4
3.3. Preciznost.....	5
3.4. AUC in ROC krivulja .....	5
3.5. F-mera .....	6
3.6. Metrika zmede .....	7
4. Analiza pridobljenih rezultatov .....	8
4.1. Točnost.....	8
4.2. Priklic.....	8
4.3. Preciznost.....	9
4.4. AUC ROC.....	9
4.5. F-Mera.....	10
4.6. Metrika zmede .....	10
5. Razmislek oz. Mnenje .....	11

## 1. Predstavitev klasifikatorja

Bayesov klasifikator je statistični klasifikator, ki omogoča izračun verjetnosti pripadnosti podatka k določenemu razredu.

Formula klasifikatorja je:

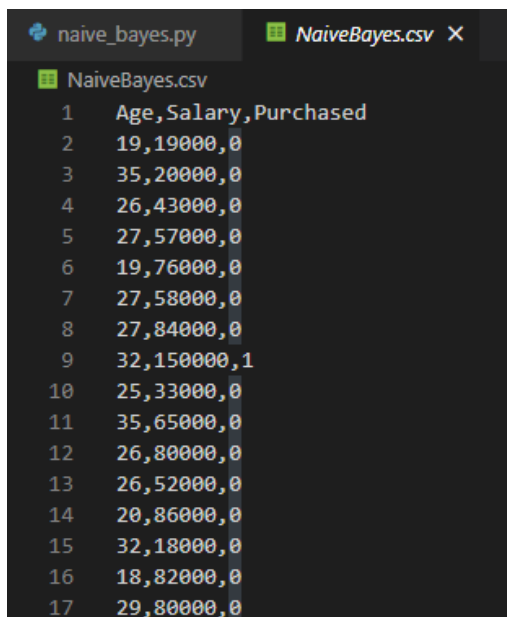
$$P(\text{class} | \text{data}) = (P(\text{data} | \text{class}) * P(\text{class})) / P(\text{data})$$

Naivni bayesov klasifikator predpostavlja, da so podatki neodvisni drug od drugega in preko neodvisnosti sklepa pripadnost podatka k razredu.

## 2. Predstavitev podatkovne zbirke

Podatkovna zbirka ima 3 kategorije (Age, Salary, Purchased) –

Glavna kategorija je Purchased (1 == Kupljeno, 0 == Ni Kupljeno), saj jo želimo napovedati na podlagi Age in Salary.



	Age	Salary	Purchased
1	Age	Salary	Purchased
2	19	19000	0
3	35	20000	0
4	26	43000	0
5	27	57000	0
6	19	76000	0
7	27	58000	0
8	27	84000	0
9	32	150000	1
10	25	33000	0
11	35	65000	0
12	26	80000	0
13	26	52000	0
14	20	86000	0
15	32	18000	0
16	18	82000	0
17	29	80000	0

### 3. Ovrednoteni rezultati klasifikacijskih algoritmov

Legenda uporabljenih treminov (klasifikatorjev):

- NN – Neural Network
- RF – Random Forest
- NB - Naivni bayes

#### 3.1. Točnost

Točnost pove koliko izračunanih podatkov je algoritem pravilno napovedal.

NB:

```
-----  
Accuracy for each fold:  
[86.25, 91.25, 90.0, 95.0, 88.75]  
Average Accuracy:  
90.25  
-----
```

Orange3:

- Neural Network: 90.5
- Random Forest: 88.7
- SVM: 90.7

#### 3.2. Priklic

Priklic pove razmerje med  $tp / (tp + fn)$  ->  $tp$ (true positive),  $fn$ (false negative) – priklic je zmožnost, da klasifikator poišče vse pozitivne primere.

Najboljša vrednost je 1, najslabša pa 0.

NB:

```
-----  
Recall for each fold:  
[0.8260869565217391, 0.92, 0.896551724137931, 1.0, 0.9393939393939394]  
Average Recall:  
0.916406524010722  
-----
```

Orange3:

- NN: 0.905
- RF: 0.887
- SVM: 0.907

### 3.3. Preciznost

Preciznost je razmerje med  $tp / (tp + fp)$  ->  $tp$ (true positive),  $fp$ (false positive) – zmožnost klasifikatorja, da pozitivnega primera ne ovrednoti kot negativnega.

NB:

```
-----
Precision for each fold:
[0.7307692307692307, 0.8214285714285714, 0.8387096774193549, 0.8918918918918919, 0.8157894736842105]
Average Precision:
0.8197177690386519
-----
```

Orange3:

- NN: 0.906
- RF: 0.888
- SVM: 0.910

### 3.4. AUC in ROC krivulja

AUC – Area under the ROC Curve: Izračun območja pod ROC krivuljo.

ROC – Receiver Operating Characteristic: metrika oceni kvaliteto izhodnih podatkov.

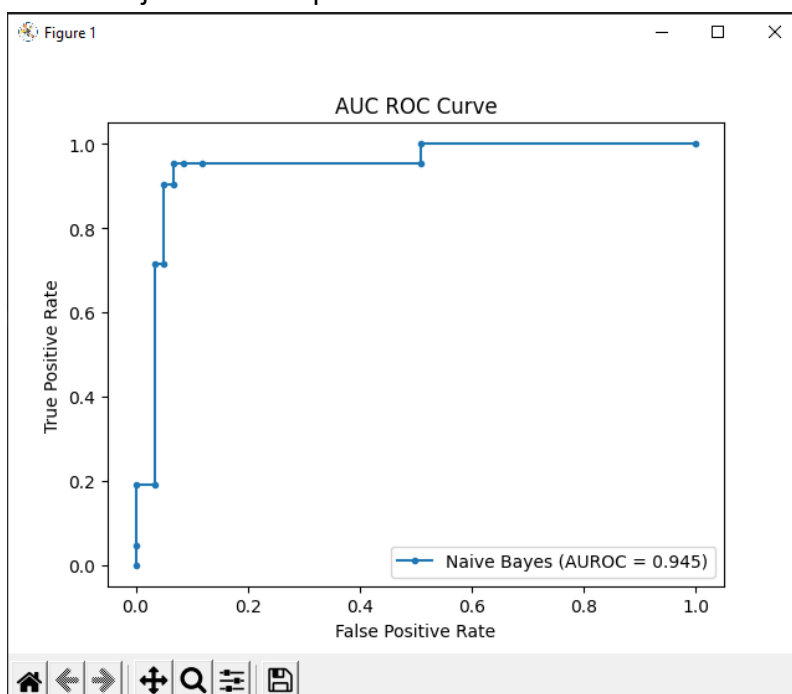
ROC AUC score: izračuna območje pod ROC krivuljo glede na napoved rezultatov.

ROC AUC score za vsaki fold in povprečje:

NB:

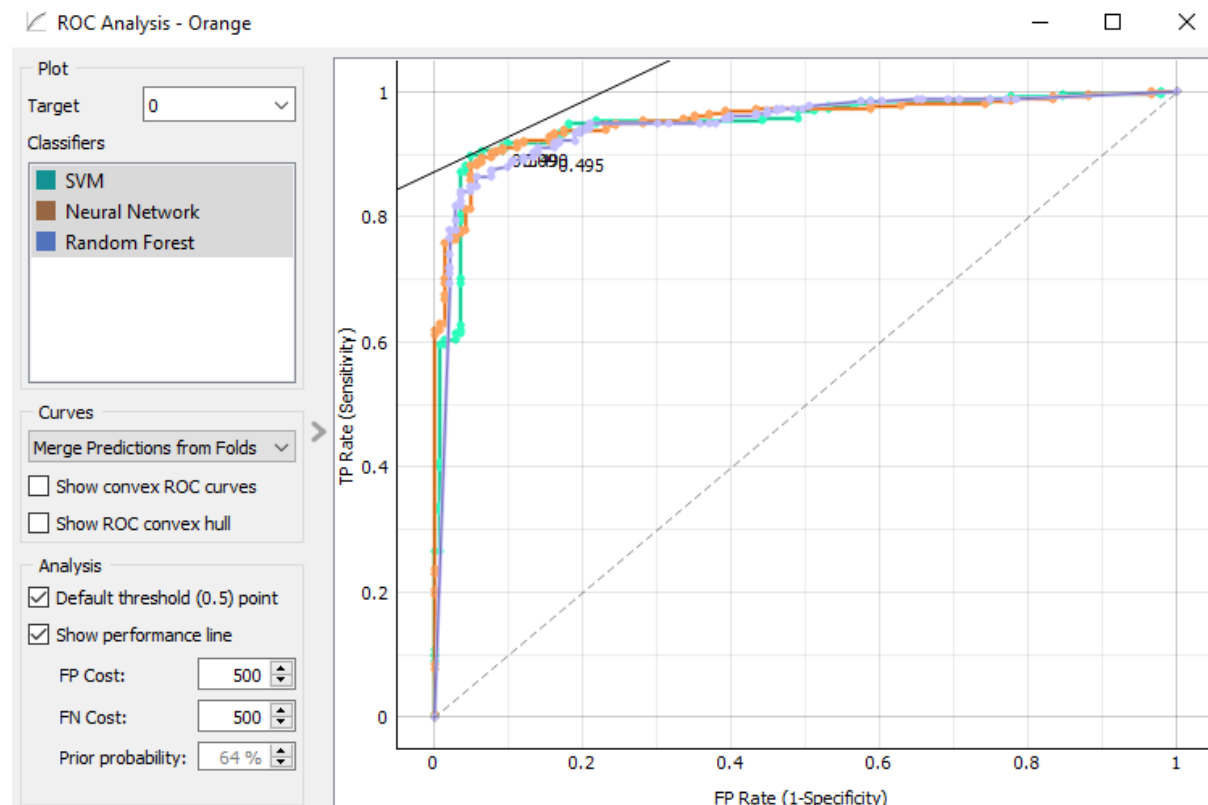
```
-----
ROC AUC for each fold:
[0.8516399694889398, 0.9145454545454546, 0.8992562542258282, 0.9574468085106382, 0.8952288845905868]
Average ROC AUC:
0.9036234742722895
-----
```

ROC krivulja za celotno podatkovno zbirko:



Orange3:

- NN: 0.955
- RF: 0.947
- SVM: 0.950



### 3.5. F-mera

F-mera je harmonično povprečje med preciznostjo in priklicem.

Formula za f-mero:  $F1 = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$

NB:

```
-----  
F1 for each fold:  
[0.7755102040816326, 0.8679245283018867, 0.8666666666666666, 0.9428571428571428, 0.8732394366197183]  
Average F1:  
0.8652395957054093  
-----
```

Orange3:

- NN: 0.905
- RF: 0.888
- SVM: 0.908

### 3.6. Metrika zmede

Metrika zmede prikaže 2x2 tabelo v kateri se izpišejo podatki v TP(true positive), FN(false negative), FP(false positive), TN(true negative)

Naivni bayes:

```
-----
Confusion for each fold:
[array([[50,  7],
       [ 4, 19]], dtype=int64), array([[50,  5],
       [ 2, 23]], dtype=int64), array([[46,  5],
       [ 3, 26]], dtype=int64), array([[43,  4],
       [ 0, 33]], dtype=int64), array([[40,  7],
       [ 2, 31]], dtype=int64)]
Average Confusion:
[[45.8  5.6]
 [ 2.2 26.4]]
-----
```

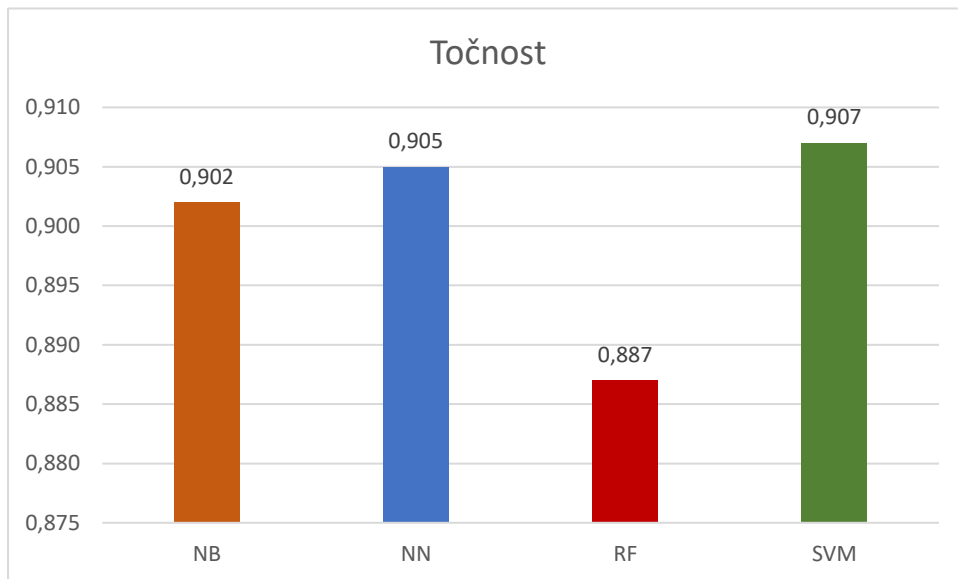
Orange3:

- NN: [235,22][16,127]
- RF: [234,23][22,121]
- SVM: [233,24][13,130]

Orange3 ima večje številke, ker v mojem primeru uporabljam podatke iz individualnega folda, torej, mojih podatkov v vsakem foldu je 80.

## 4. Analiza pridobljenih rezultatov

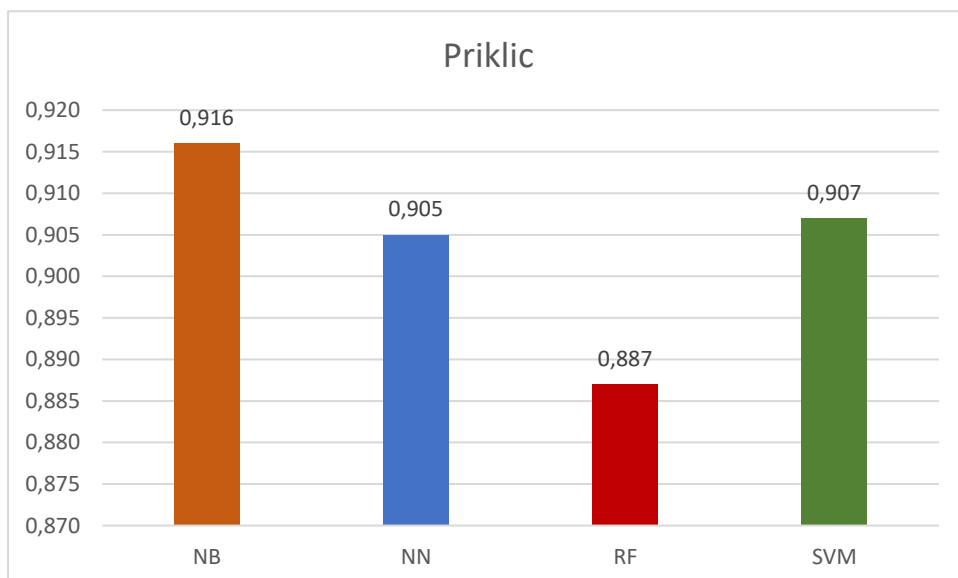
### 4.1. Točnost



NB (0,902) je v primerjavi z NN (0,905) in SVM (0,907) klasifikatorjema dokaj enak, razlika med njimi je: (NN 0,003 +) (SVM 0,005 +)

Vidno slabši pa je RF klasifikator, ki ima naj-nižjo točnost 0,887 oz. 88,7%

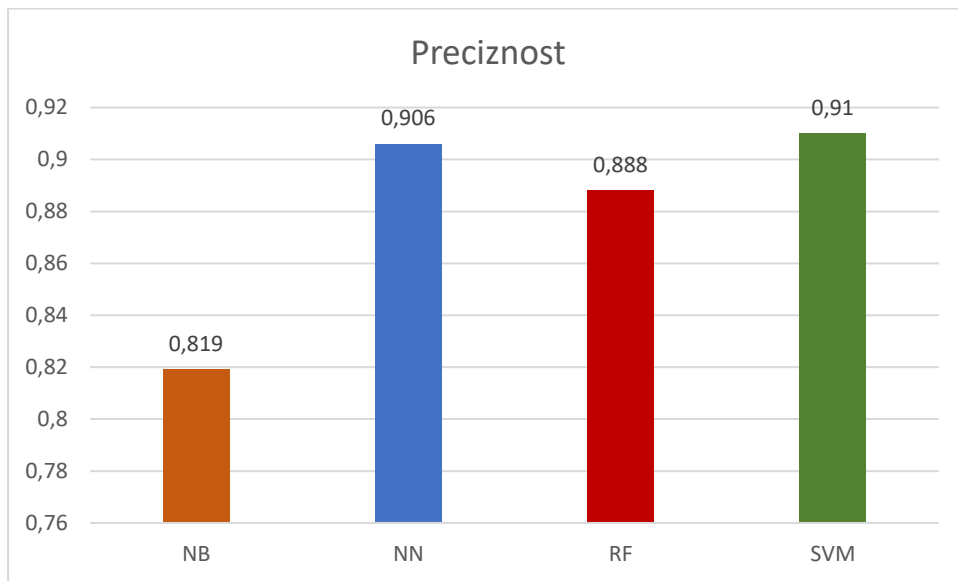
### 4.2. Priklic



NB (0,916) ima najboljši rezultat priklica. Drugi je SVM (0,907), tretji NN (0,905) in zadnji RF (0,887). RF ima ponovno najslabši rezultat. NN in SVM sta zelo blizu, na grafu izstopa NB kot najboljši.

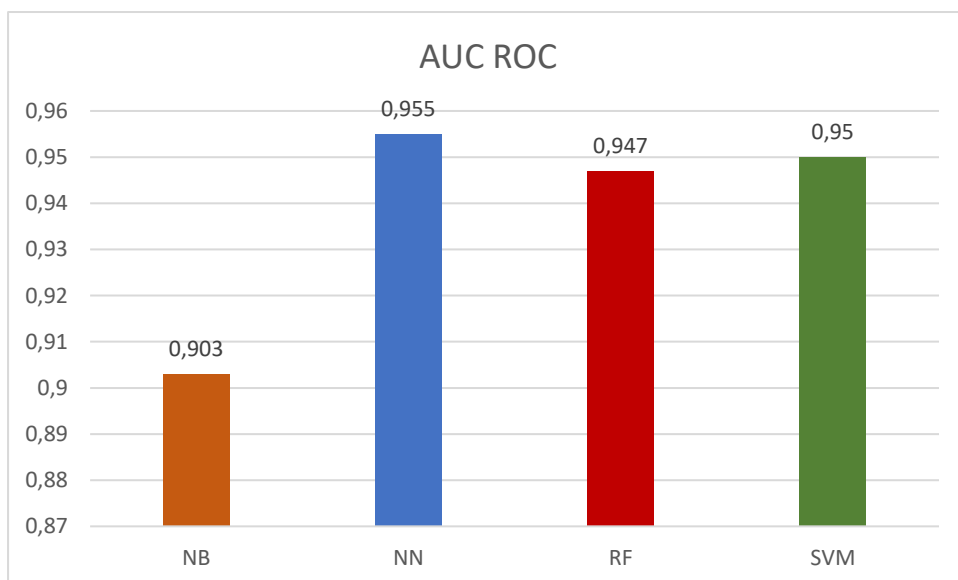


### 4.3. Preciznost



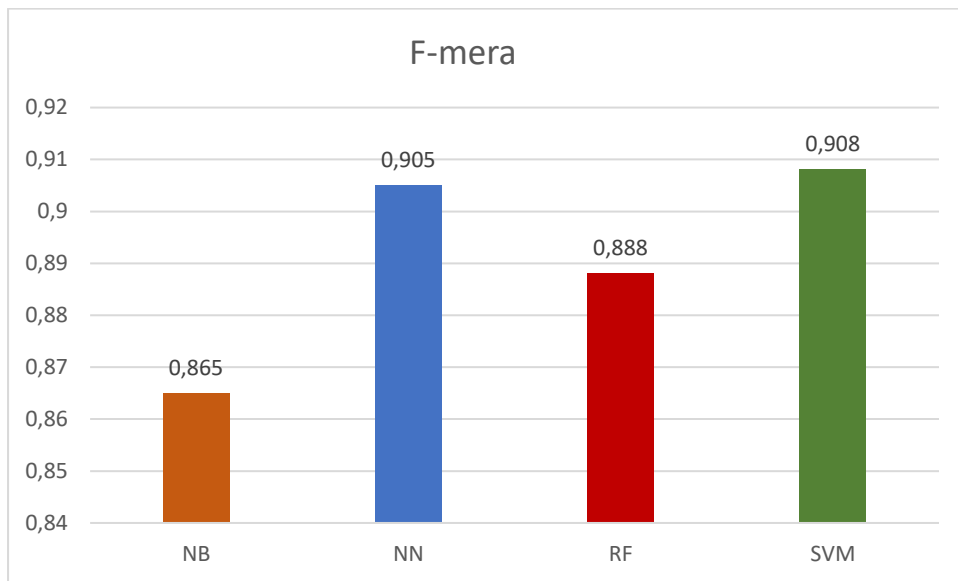
NB (0,819) je pri preciznosti najslabši, medtem sta NN (0,906) in SVM (0,91) blizu, od njiju slabši pa je RF (0,888).

### 4.4. AUC ROC



NB (0,903) je ponovno najslabši, tokrat pri AUC ROC rezultatu. Ostali klasifikatorji NN (0,955), RF (0,947) in SVM (0,95) so dokaj blizu, najboljši je NN, najslabši med tremi pa je RF.

#### 4.5. F-Mera



Pri F-meri je najboljši SVM (0,908), drugi NN (0,905), tretji RF (0,888) zadnji pa NB (0,865).

#### 4.6. Metrika zmede

**NN**

	0	1	$\Sigma$
0	235	22	257
1	16	127	143
$\Sigma$	251	149	400

**RF**

	0	1	$\Sigma$
0	234	23	257
1	20	123	143
$\Sigma$	254	146	400

**SVM**

	0	1	$\Sigma$
0	233	24	257
1	13	130	143
$\Sigma$	246	154	400

**NB**

	0	1
0	45.8	5.6
1	2.2	26.4

NB je povprečje vseh 5 foldov in v vsakem foldu je 80 podatkov, zato so številke manjše kot podatki iz Orange3.

Procentualno gledano:

- True positive NN (58,75%), RF (58,5%), SVM (58,25%), NB (57,25%) – najvišji NN
- False negative NN (5,5%), RF (5,75%), SMV (6%), NB (7%) – najvišji NB
- False positive NN (4%), RF (5%), SMV (3,25%), NB (2,75%) – najvišji RF
- True negative NN (31,75%), RF (30,75%), SVM (32,5%), NB (33%) – najvišji NB

TP – vsi klasifikatorji so blizu, NB je najslabši. TN – NB ima tukaj najvišji procent, najmanjšega pa NN. FP – najvišji procent ima RF, najmanjši pa je pri NB. FN – najvišji je NB, najnižji pa RF.

V vseh kategorijah so klasifikatorji blizu en drugemu, največje deviacije so pri FN med NB in RF, pri čem je višji NB, razlikujeta se za 2,25. Za enako število se razlikujeta tudi RF in NB pri FP, kjer je najvišji RF.

## 5. Razmislek oz. Mnenje

Neural networks in Random forest sta oba zelo močna in natančna algoritma, kadar je vključeno veliko podatkov. Podatkov pa v tem primeru ni veliko, zato je RF vedno zadnji ali predzadnji pri vseh metrikah. Naivni bayes se je od vseh odrezal najboljše pri priklicu, saj deluje boljše pri majhnemu številu podatkov in kadar so atributi neodvisni en od drugega (to omogoča lažje iskanje pozitivnih podatkov). SVM deluje boljše od Naivnega bayesa, kadar so atributi odvisni en od drugega – v tem primeru so odvisni, saj je SVM vedno prvi ali drugi. Vsi klasifikatorji so boljši od NB v preciznost, saj so bolje optimizirani za natančnost, ko se atributi med seboj povezujejo. Pri izračunu ROC AUC so boljši naprednejši algoritmi RF, NN in SVM, ker so bolj natančni, ter tako bolje napovejo rezultate – iz tega sledi, da je ROC AUC score boljši. Algoritmi so bili boljši od NB v preciznosti in posledično tudi boljši v F-meri, saj je ta odvisna od razmerja med preciznostjo in priklicem. Pri metriki zmede se je procentualno gledano najboljše odnesel SVM, saj lahko napove največ resnično pozitivnih in resnično negativnih števil – spet zaradi tega, ker se bolje odnese, ko se podatki med seboj povezujejo in ko jih je manj.

NN in RF bi potrebovala večjo količino podatkov, da bi se njuna moč in natančnost bolj izkazala.