Uticaj metoda imputacije podataka na tačnost predviđanja

Fakultet organizacionih nauka

Mentor:
dr Bratislav Petrović

UNIVERZITET U BEOGRADU
FAKULTET ORGANIZACIONIH NAUKA

Student:
Mihailo Stupar

Agenda

- Formulacija problema
- Imputacija podataka
- Eksperiment
- Predložena metoda imputacije
- Zaključak

Formulacija problema

Predviđanje nad kompletnim skupom podataka

$$Y = f(X_1, X_2, X_3)$$

X_1	X_2	X_3	Y
3	104	1	34
5	98	1	41
3	112	2	23
4	142	1	45
6	87	2	42
5	99	2	23
1	102	1	54
3	88	1	37
5	92	1	32
7	108	2	42
2	79	2	34
4	91	1	39

Formulacija problema

Predviđanje nad kompletnim skupom podataka

$Y = f(X_1, X_2, X_3)$

Predviđanje nad skupom podataka sa nedostajućim vrednostima

Y = ?

X_1	X_2	X_3	Y
	104	1	34
5			41
	112	2	23
	142		45
6		2	42
	99	2	23
1			54
3	88	1	37
	92	1	32
7		2	42
2	79		34
	91	1	39

Imputacija podataka

X_1	X_2	X_3	Y
	104	1	34
5			41
	112	2	23
	142		45
6		2	42
	99	2	23
1			54
3	88	1	37
	92	1	32
7		2	42
2	79		34
	91	1	39

Imputacija podataka

$$X_1 = f(Y, X_2, X_3)$$

X_1	X_2	X_3	Y
	104	1	34
5			41
	112	2	23
	142		45
6		2	42
	99	2	23
1			54
3	88	1	37
	92	1	32
7		2	42
2	79		34
	91	1	39

Imputacija podataka

$X_1 = f(Y, X_2, X_3)$

$X_2 = f(X_1, Y, X_3)$

X_1	X_2	X_3	Y
	104	1	34
5			41
	112	2	23
	142		45
6		2	42
	99	2	23
1			54
3	88	1	37
	92	1	32
7		2	42
2	79		34
	91	1	39

Imputacija podataka

$X_1 = f(Y, X_2, X_3)$

$X_2 = f(X_1, Y, X_3)$

$X_3 = f(X_1, X_2, Y)$

X_1	X_2	X_3	Y
	104	1	34
5			41
	112	2	23
	142		45
6		2	42
	99	2	23
1			54
3	88	1	37
	92	1	32
7		2	42
2	79		34
	91	1	39

Imputacija podataka

$X_1 = f(Y, X_2, X_3)$

$X_2 = f(X_1, Y, X_3)$

$X_3 = f(X_1, X_2, Y)$

f **–** linearna regresija

f **–** stohastička linearna regresija

f **–** šuma stabala odlučivanja

X_1	X_2	X_3	Y
	104	1	34
5			41
	112	2	23
	142		45
6		2	42
	99	2	23
1			54
3	88	1	37
	92	1	32
7		2	42
2	79		34
	91	1	39

Mere efikasnosti (tačnosti) imputacije

Srednja kvadratna greška

X_1	X_2	X_3	Y
5 (3)	104	1	34
5	101 (98)	1 (1)	41
4 (3)	112	2	23
2 (4)	142	2 (1)	45
6	87 (87)	2	42
5 (5)	99	2	23
1	98 (102)	1.6 (1)	54
3	88	1	37
3 (5)	92	1	32
7	99 (108)	2	42
2	79	3 (2)	34
3 (4)	91	1	39

Mere efikasnosti (tačnosti) imputacije

Srednja kvadratna greška

Koren srednje kvadratne greške

X_1	X_2	X_3	Y
5 (3)	104	1	34
5	101 (98)	1 (1)	41
4 (3)	112	2	23
2 (4)	142	2 (1)	45
6	87 (87)	2	42
5 (5)	99	2	23
1	98 (102)	1.6 (1)	54
3	88	1	37
3 (5)	92	1	32
7	99 (108)	2	42
2	79	3 (2)	34
3 (4)	91	1	39

Mere efikasnosti (tačnosti) imputacije

Srednja kvadratna greška

Koren srednje kvadratne greške

Prosečna relativna greška

X_1	X_2	X_3	Y
5 (3)	104	1	34
5	101 (98)	1 (1)	41
4 (3)	112	2	23
2 (4)	142	2 (1)	45
6	87 (87)	2	42
5 (5)	99	2	23
1	98 (102)	1.6 (1)	54
3	88	1	37
3 (5)	92	1	32
7	99 (108)	2	42
2	79	3 (2)	34
3 (4)	91	1	39

Mere efikasnosti (tačnosti) imputacije

Srednja kvadratna greška

Koren srednje kvadratne greške

Prosečna relativna greška

Koren srednje kvadratne greške prilikom predviđanja linearnom regresijom

X_1	X_2	X_3	Y
5	104	1	34
5	101	1	41
4	112	2	23
2	142	2	45
6	87	2	42
5	99	2	23
1	98	1.6	54
3	88	1	37
3	92	1	32
7	99	2	42
2	79	3	34
3	91	1	39

Eksperiment

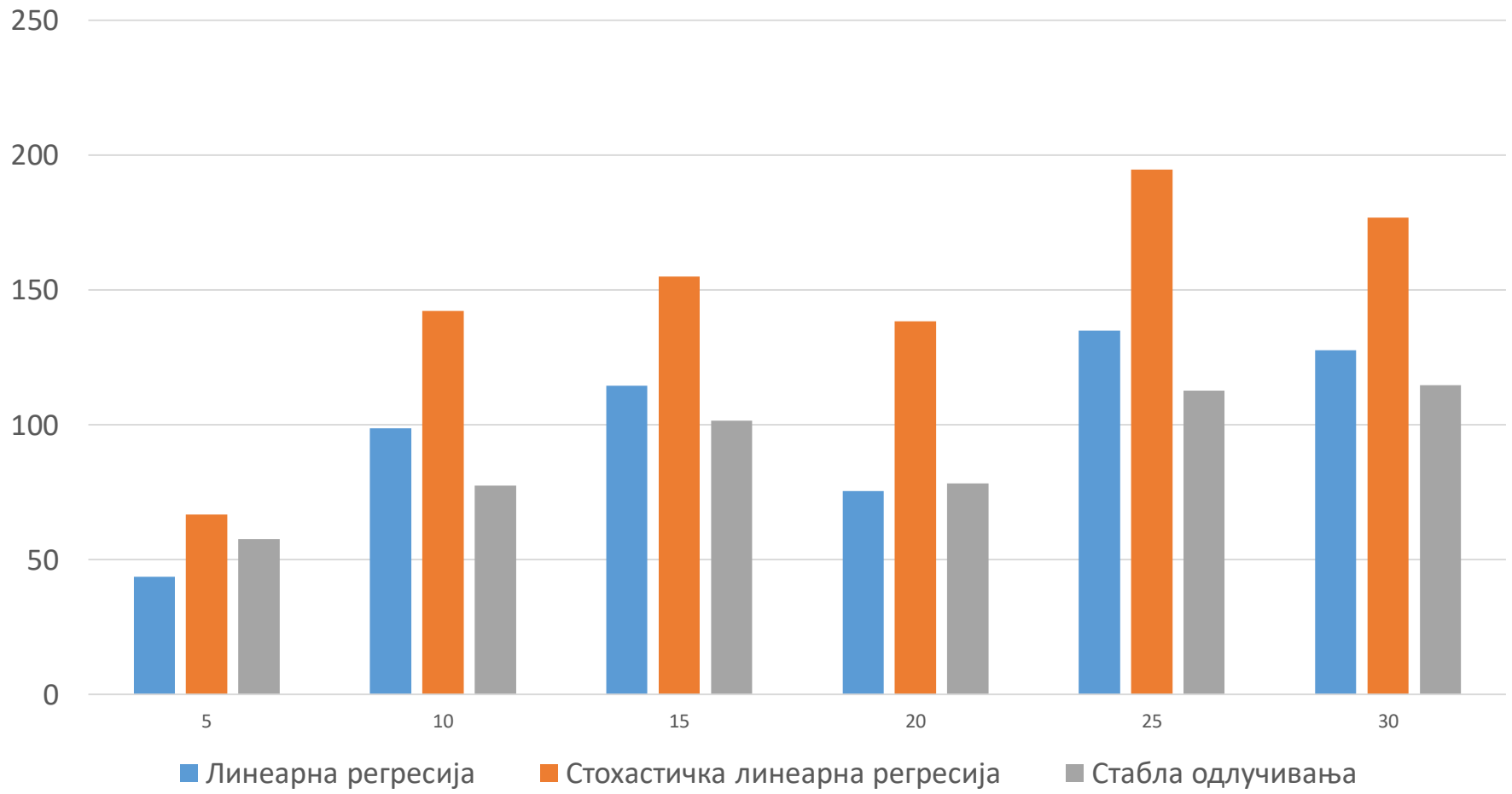
6 skupova podataka (5% , 10% , 15% , 20% , 25% , 30%)

Imputacija linearnom regresijom – *R* jezik, *mice* biblioteka

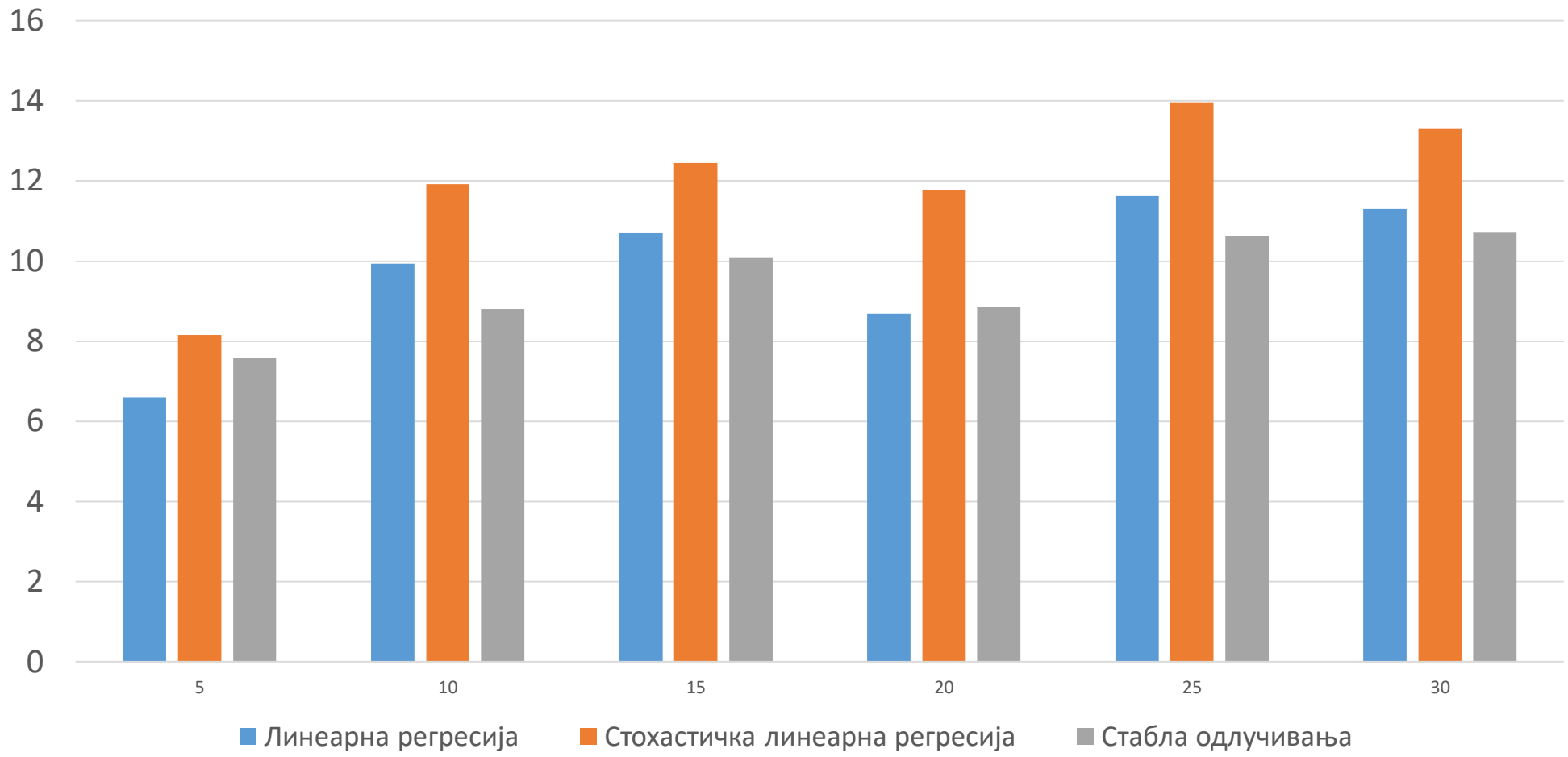
Imputacija stohastičkom linearnom regresijom – *R* jezik, *mice* biblioteka

Imputacija šumom stabala odlučivanja – *R* jezik, *missForest* biblioteka

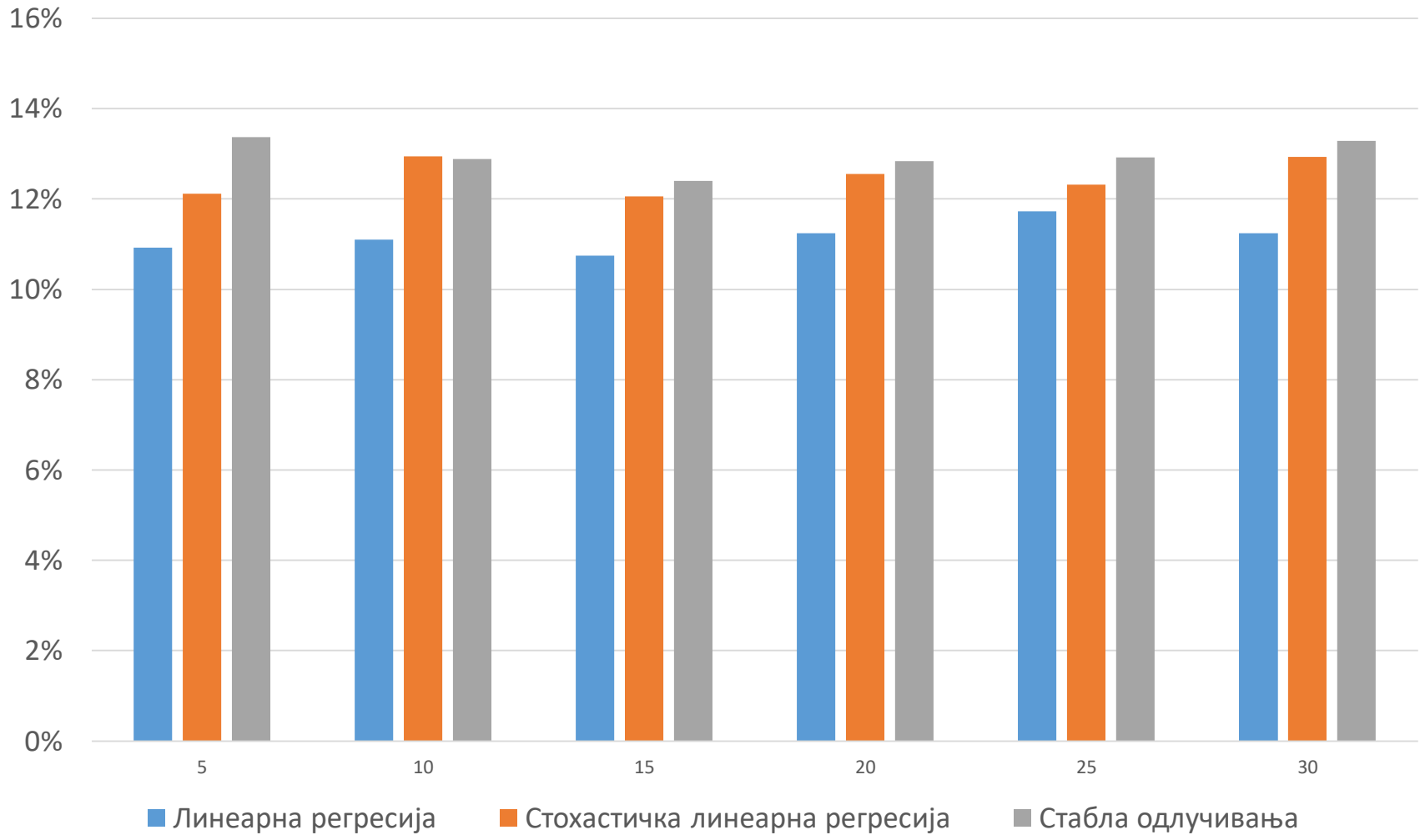
Rezultati – Srednja kvadratna greška



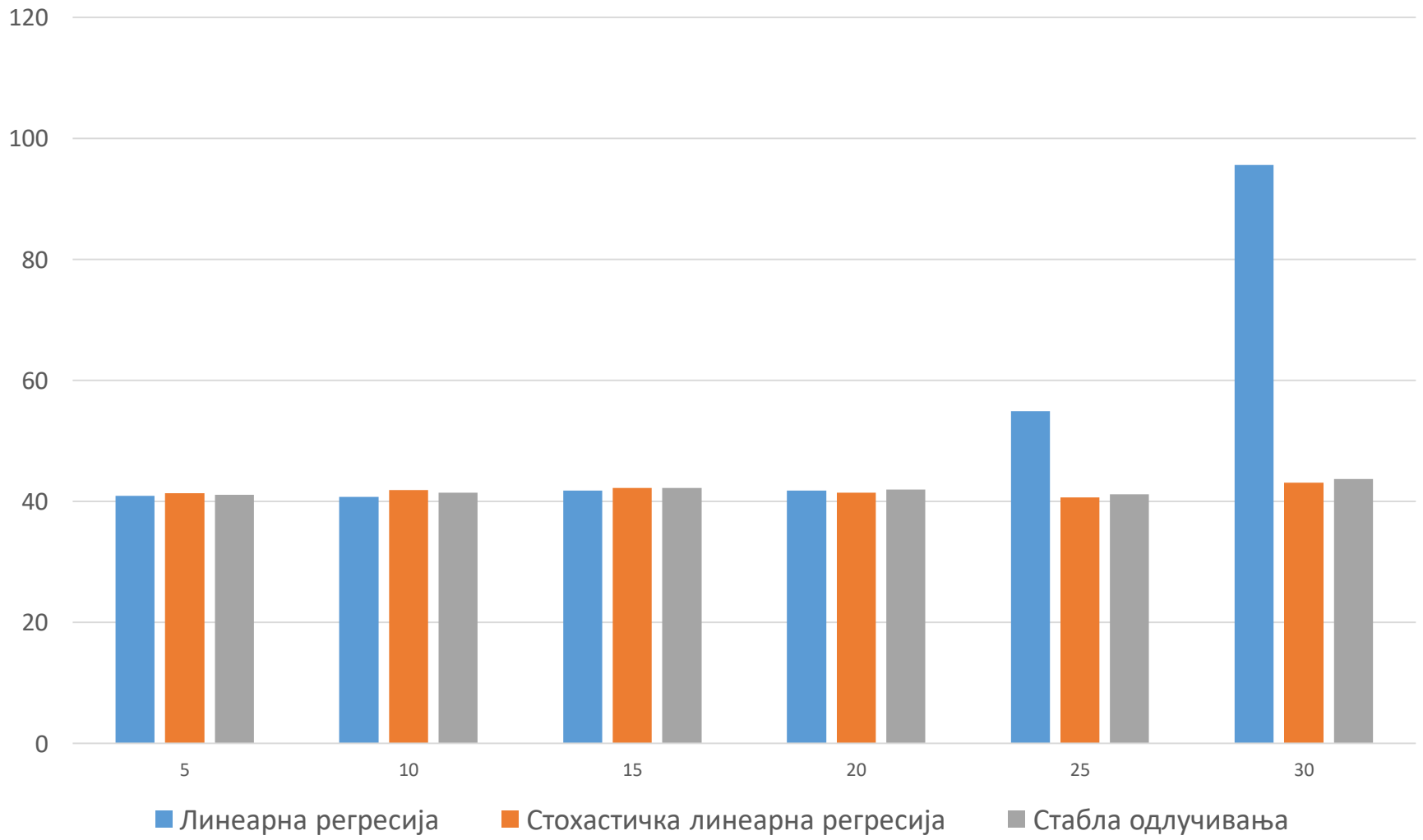
Rezultati – Koren srednje kvadratne greške



Rezultati – Prosečna relativna greška



Rezultati – Koren srednje kvadratne greške predviđanja



Predložena metoda imputacije

X_1	X_2	X_3	Y
	104	1	34
5			41
	112	2	23
	142		45
6		2	42
	99	2	23
1			54
3	88	1	37
	92	1	32
7		2	42
2	79		34
	91	1	39
4			44
	88		43
2	101	2	39
3		2	29

Predložena metoda imputacije

X_1	X_2	X_3	Y
	104	1	34
5			41
	112	2	23
	142		45
6		2	42
	99	2	23
1			54
3	88	1	37
	92	1	32
7		2	42
2	79		34
	91	1	39
4			44
	88		43
2	101	2	39
3		2	29

K-srednjih vrednosti klasterovanje

X_1	X_2	X_3	Y
	104	1	34
5			41
	112	2	23
	142		45
6		2	42

X_1	X_2	X_3	Y
	99	2	23
1			54
3	88	1	37
	92	1	32

X_1	X_2	X_3	Y
7		2	42
2	79		34
	91	1	39
4			44
	88		43
2	101	2	39
3		2	29

Predložena metoda imputacije

X_1	X_2	X_3	Y
	104	1	34
5			41
	112	2	23
	142		45
6		2	42
	99	2	23
1			54
3	88	1	37
	92	1	32
7		2	42
2	79		34
	91	1	39
4			44
	88		43
2	101	2	39
3		2	29

K-srednjih vrednosti klasterizacija

X_1	X_2	X_3	Y
	104	1	34
5			41
	112	2	23
	142		45
6		2	42

X_1	X_2	X_3	Y
	99	2	23
1			54
3	88	1	37
	92	1	32

X_1	X_2	X_3	Y
7		2	42
2	79		34
	91	1	39
4			44
	88		43
2	101	2	39
3		2	29

Imputacija stohastičkom linearnom regresijom

X_1	X_2	X_3	Y
7	104	1	34
5	101	2	41
4	112	2	23
7	142	1	45
6	131	2	42

X_1	X_2	X_3	Y
2	99	2	23
1	94	1	54
3	88	1	37
1	92	1	32

X_1	X_2	X_3	Y
7	88	2	42
2	79	2	34
5	91	1	39
4	91	2	44
5	88	1	43
2	101	2	39
3	104	2	29

Predložena metoda imputacije

X_1	X_2	X_3	Y
	104	1	34
5			41
	112	2	23
	142		45
6		2	42
	99	2	23
1			54
3	88	1	37
	92	1	32
7		2	42
2	79		34
	91	1	39
4			44
	88		43
2	101	2	39
3		2	29

K-srednjih vrednosti klasterizacija

X_1	X_2	X_3	Y
	104	1	34
5			41
	112	2	23
	142		45
6		2	42

X_1	X_2	X_3	Y
	99	2	23
1			54
3	88	1	37
	92	1	32

X_1	X_2	X_3	Y
7		2	42
2	79		34
	91	1	39
4			44
	88		43
2	101	2	39
3		2	29

Imputacija stohastičkom linearnom regresijom

X_1	X_2	X_3	Y
7	104	1	34
5	101	2	41
4	112	2	23
7	142	1	45
6	131	2	42

X_1	X_2	X_3	Y
2	99	2	23
1	94	1	54
3	88	1	37
1	92	1	32

X_1	X_2	X_3	Y
7	88	2	42
2	79	2	34
5	91	1	39
4	91	2	44
5	88	1	43
2	101	2	39
3	104	2	29

X_1	X_2	X_3	Y
7	104	1	34
5	101	2	41
4	112	2	23
7	142	1	45
6	131	2	42
2	99	2	23
1	94	1	54
3	88	1	37
1	92	1	32
7	88	2	42
2	79	2	34
5	91	1	39
4	91	2	44
5	88	1	43
2	101	2	39
3	104	2	29

K-srednjih vrednosti klasterovanje

```
ulaz: k,X

izabrati k centroida
ponavljaj:
  napravi k klastera pridružujući najbliže elemente centroidima
  promeni poziciju centroida
dok centroidi ne agregišu
```

K-srednjih vrednosti klasterovanje - rastojanje

x:[5, 2, 6, 2]
y:[2, 4, , 3]

m:4/3

K-srednjih vrednosti klasterovanje - rastojanje

```
x:[5, 2, 6, 2]
y:[2, 4,  , 3]
```

```
m:4/3
```

```
x:[5, 2, 6, 2]
y:[2, 4, 6, 3]
```

K-srednjih vrednosti klasterovanje - rastojanje

```
x:[5, 2, 6, 2]          m:4/3
y:[2, 4,  , 3]

x:[5, 2, 6, 2]
y:[2, 4, 6, 3]
```

```
  [3,-2, 0,-1] // razlika
```

K-srednjih vrednosti klasterovanje - rastojanje

```
x:[5, 2, 6, 2]          m:4/3
y:[2, 4,  , 3]

x:[5, 2, 6, 2]
y:[2, 4, 6, 3]
-----------------------
   [3,-2, 0,-1] // razlika
-----------------------
   [9, 4, 0, 1] // kvadriranje
```

K-srednjih vrednosti klasterovanje - rastojanje

```
x:[5, 2, 6, 2]          m:4/3
y:[2, 4,  , 3]

x:[5, 2, 6, 2]
y:[2, 4, 6, 3]
-----------------------
  [3,-2, 0,-1] // razlika
-----------------------
  [9, 4, 0, 1] // kvadriranje
-----------------------
  14           // suma
```

K-srednjih vrednosti klasterovanje - rastojanje

```
x:[5, 2, 6, 2]          m:4/3
y:[2, 4,  , 3]

x:[5, 2, 6, 2]
y:[2, 4, 6, 3]
```

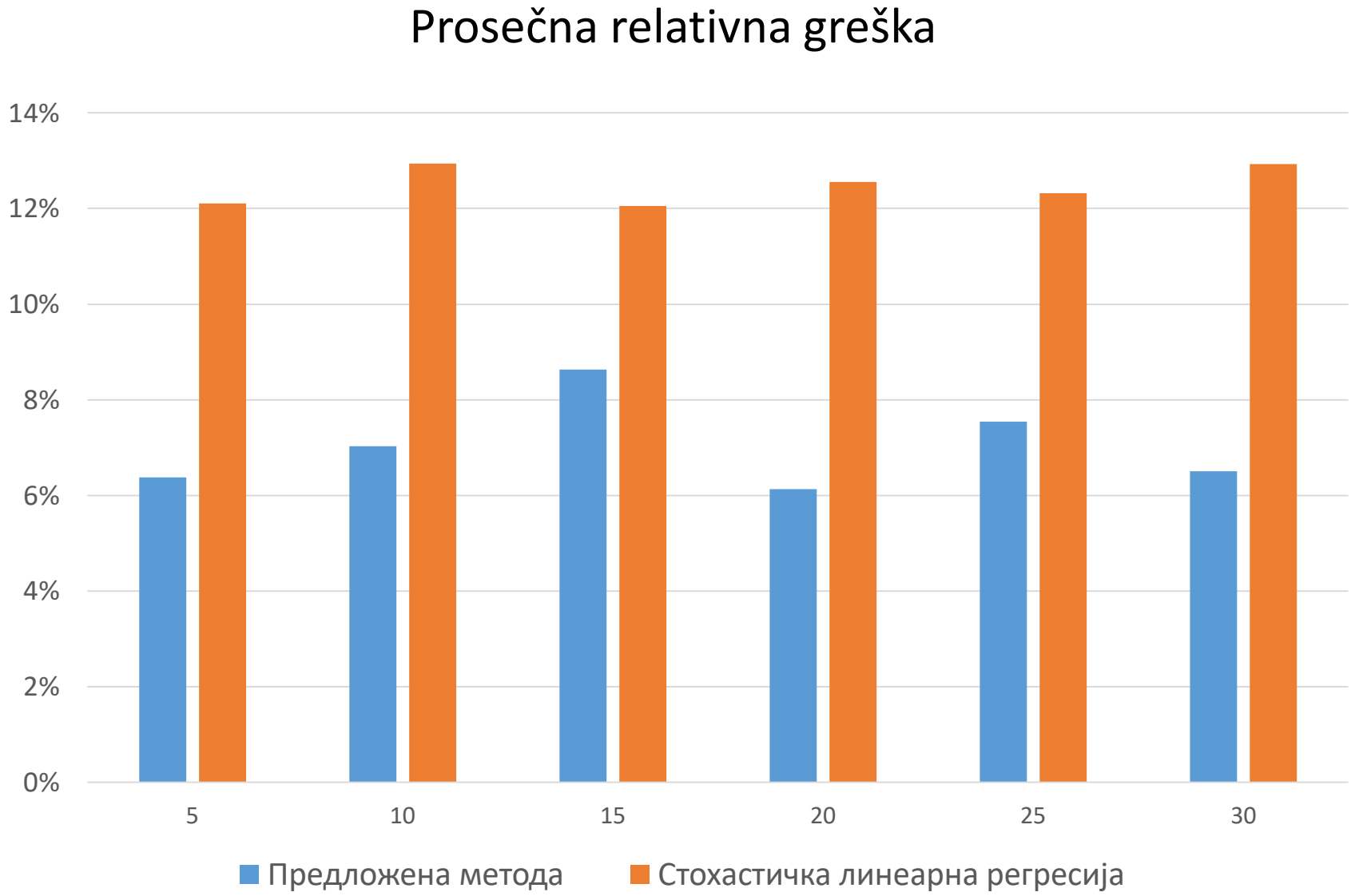
```
   [3,-2, 0,-1] // razlika
```

```
   [9, 4, 0, 1] // kvadriranje
```

```
   14           // suma
```

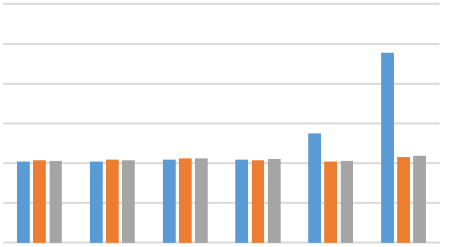
$$d: \sqrt{14 * 4/3}$$

Rezultati dobijeni preloženom metodom

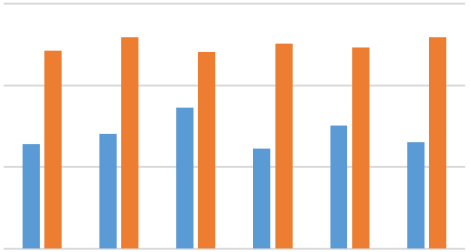


Zaključak

Imputacija podataka ima uticaj na tačnost predviđanja



Priprema podataka unapređuje tačnost imputacije



Ideja za dalje istraživanje:

PCA kao još jedan korak pripreme pred imputaciju