Klasikacija podataka prikupljenih u okviru eksperimenta o kratkim sastancima na slepo

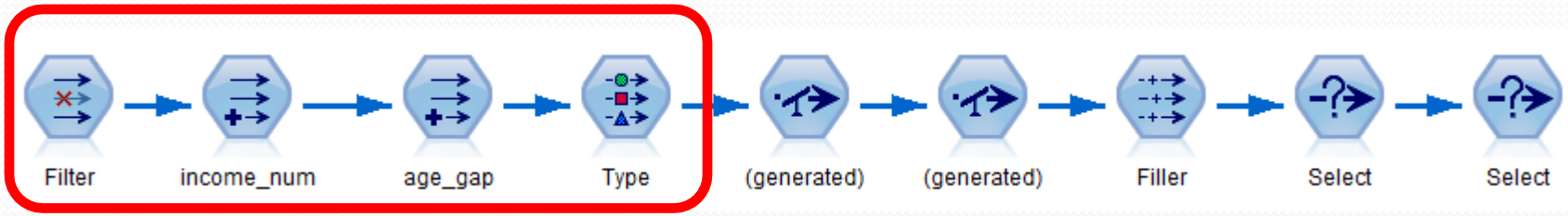
Darinka Zobenica

Eksperiment

- rađen u Kolumbiji
- sastanci od 4 minuta
- 9-21 sastanaka po osobi
- 392 ispitanika, ukupno 8378 sastanaka
- 190 kolona sa prikupljenim podacima o svakom sastanku, uključujući podatke o oba ispitanika, međusobnom sviđanju, i kasnijim interakcijama

Priprema podataka - Odabir

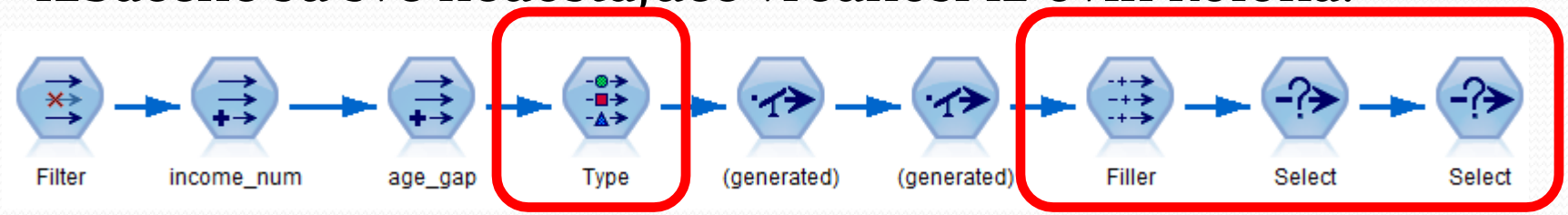
- match
- wave
  - Izbačen 12. zbog specifičnosti u okviru eksperimenta.
- int_corr
- age, age_o
- age_gap
- income_num
  - Izvršena konverzija niske u numerički tip.



Priprema podataka - Odabir

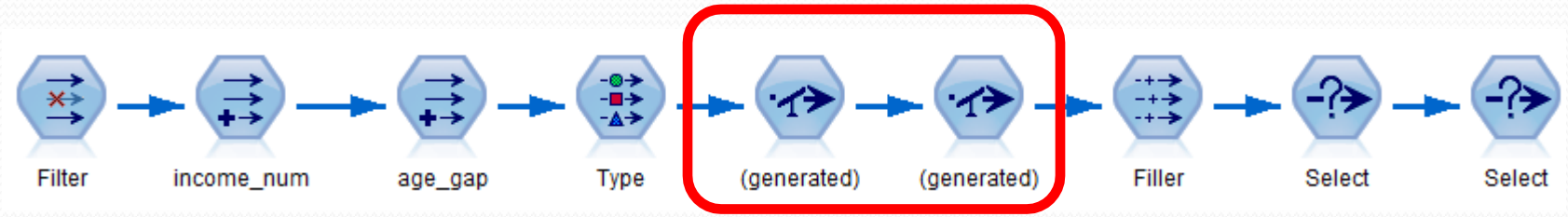
- race, race_o:
  1. Negroidna rasa
  2. Evropeidna rasa
  3. Hispano i Latinoamerikanci
  4. Mongoloidna rasa
  5. Americki starosedeoci
  6. Ostali
- samerace
- imprace

Izbačene su sve nedostajuće vrednosi iz ovih kolona.

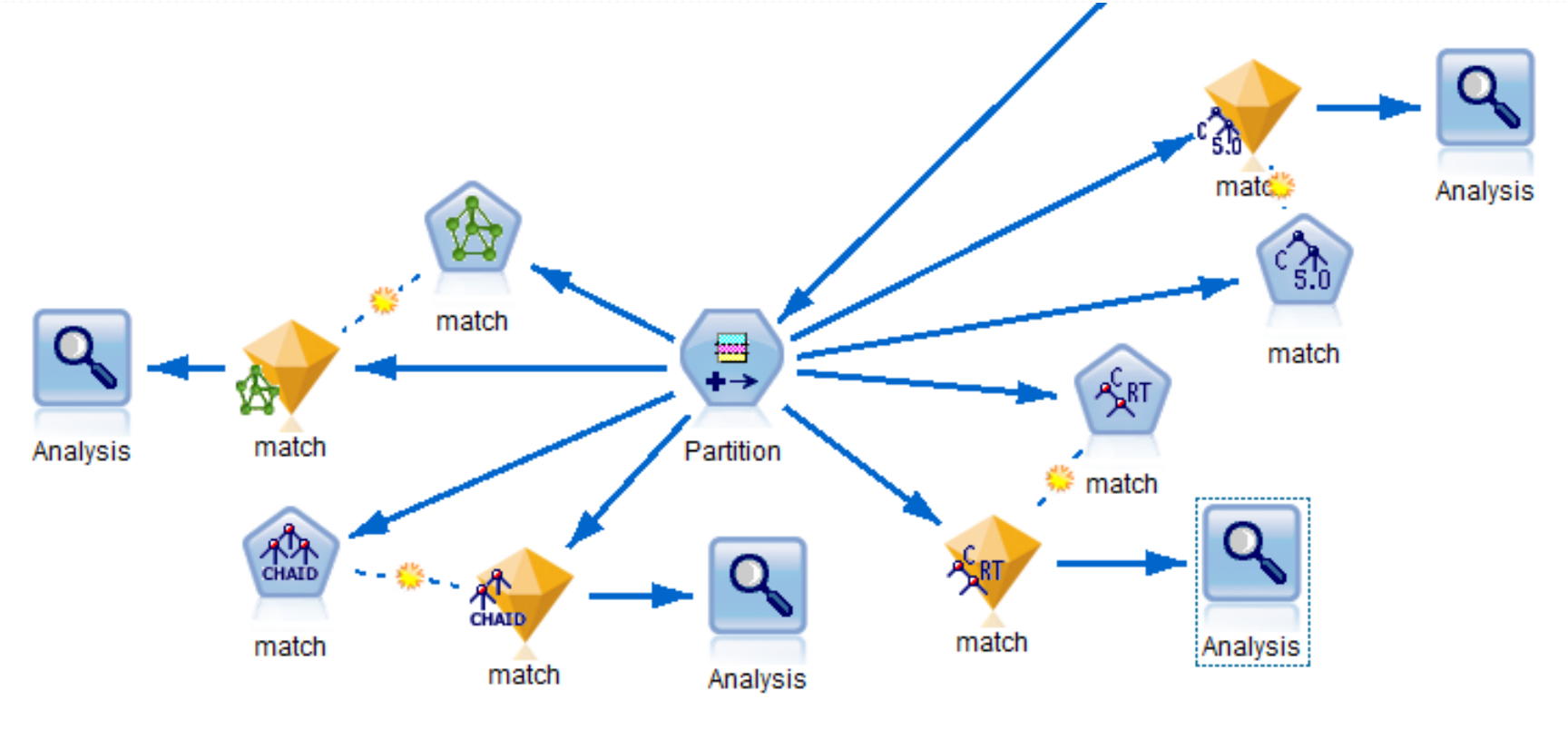


Priprema podataka - Balansiranje

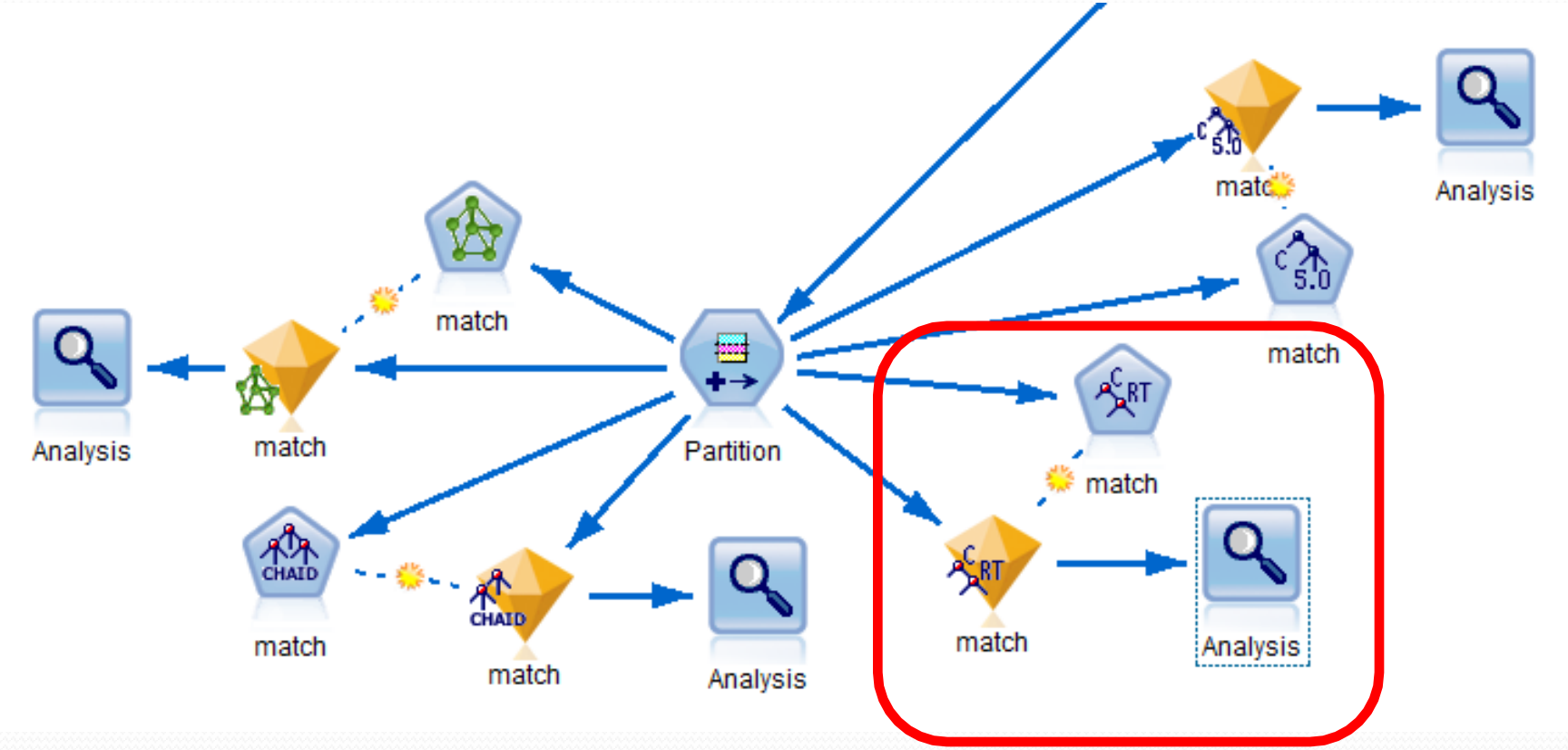
- Balansiranje može dovesti do pristrasnosti podataka.
- Izbalansirana je samo rasa, zbog većinske zastupljenosti belaca u uzorku.



Klasifikacija



C&R Tree


Analysis
match
match
Partition
CHAID
match
CHAID
match
Analysis
match
match
match
Analysis
match
Analysis

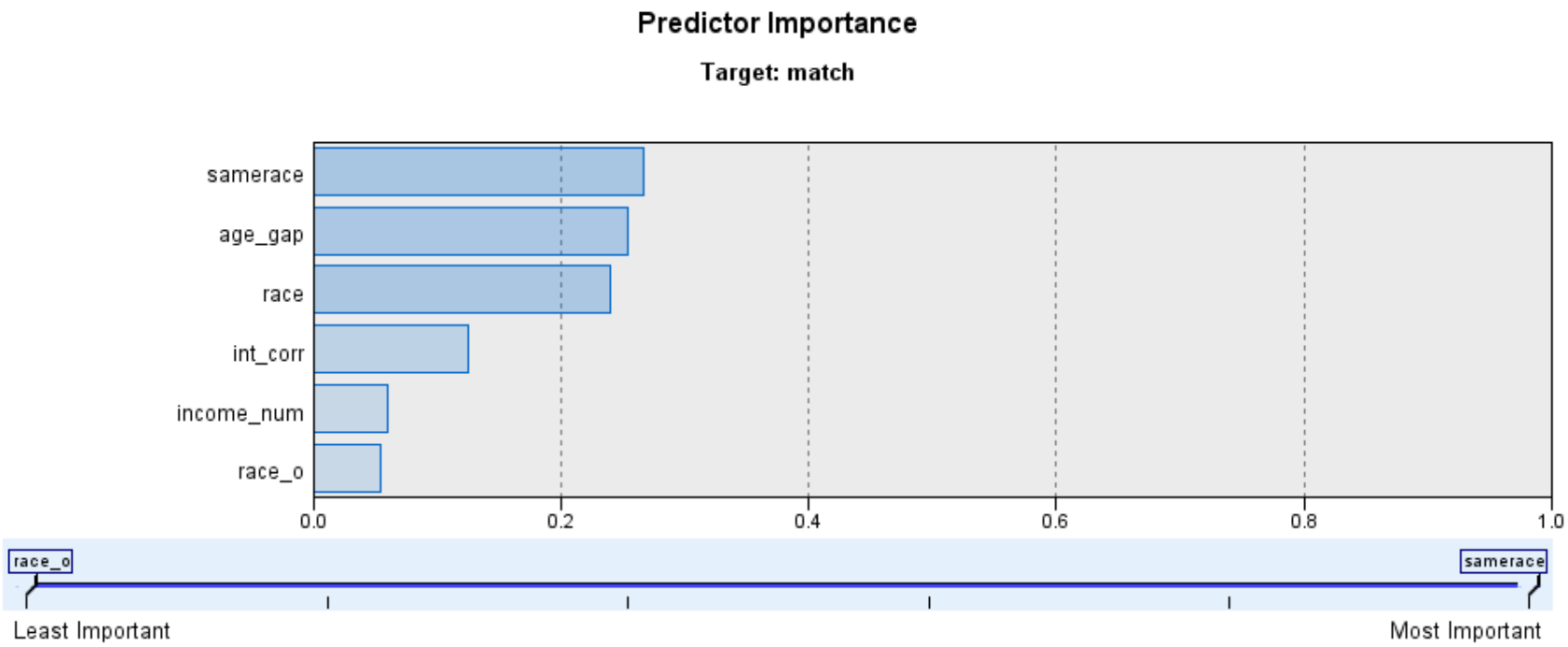
C&R Tree

Results for output field match

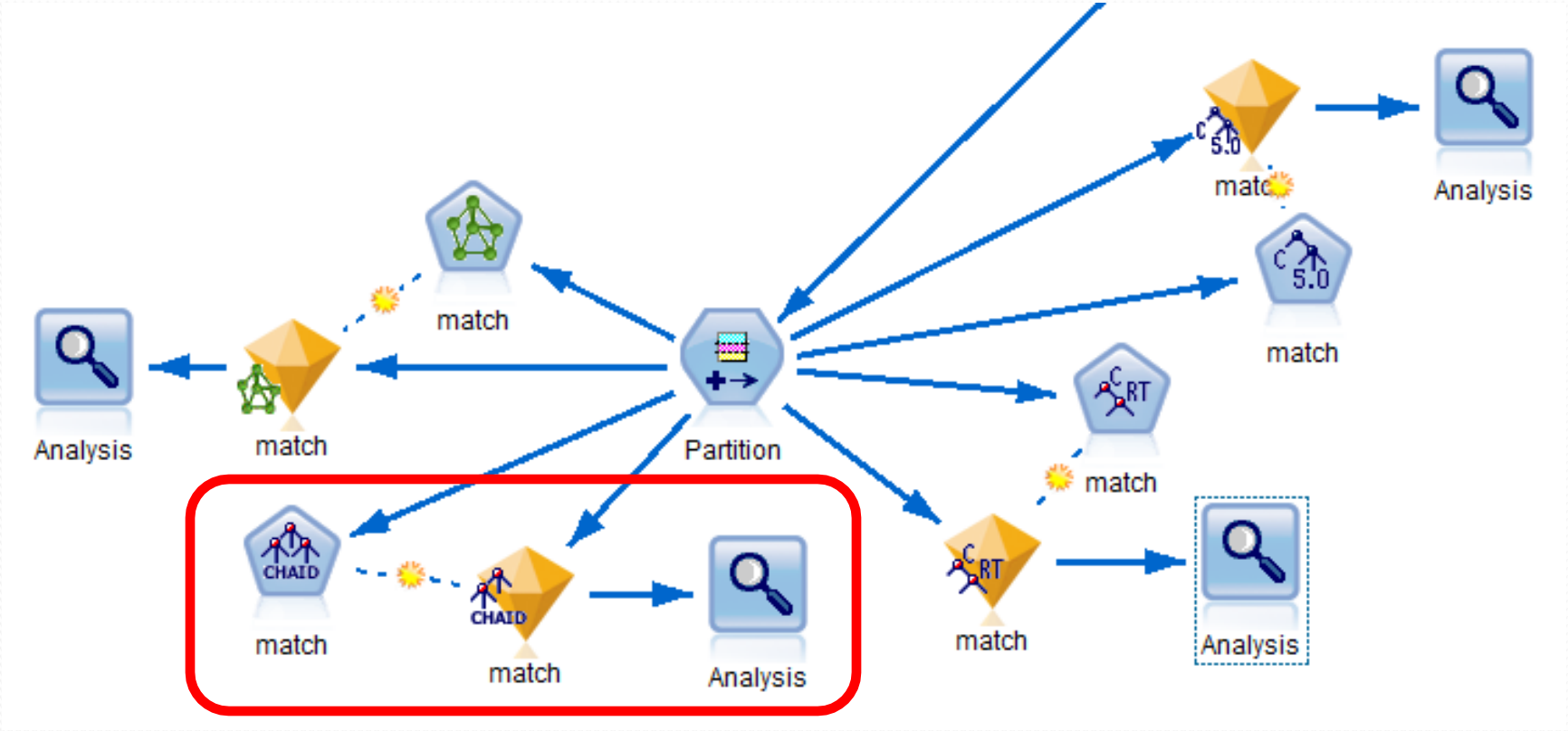
Comparing $R-match with match

'Partition'	1_Training		2_Testing	
Correct	15,005	81.24%	6,444	81.6%
Wrong	3,464	18.76%	1,453	18.4%
Total	18,469		7,897	

C&R Tree


Predictor Importance
Target: match
samerace
age_gap
race
int_corr
income_num
race_o
0.0
0.2
0.4
0.6
0.8
1.0
race_o
samerace
Least Important
Most Important

CHAID


Analysis
match
match
Partition
match
match
match
match
Analysis
match
match
Analysis
match
Analysis
CHAID
CHAID
C5.0
C5.0
C&RT
C&RT

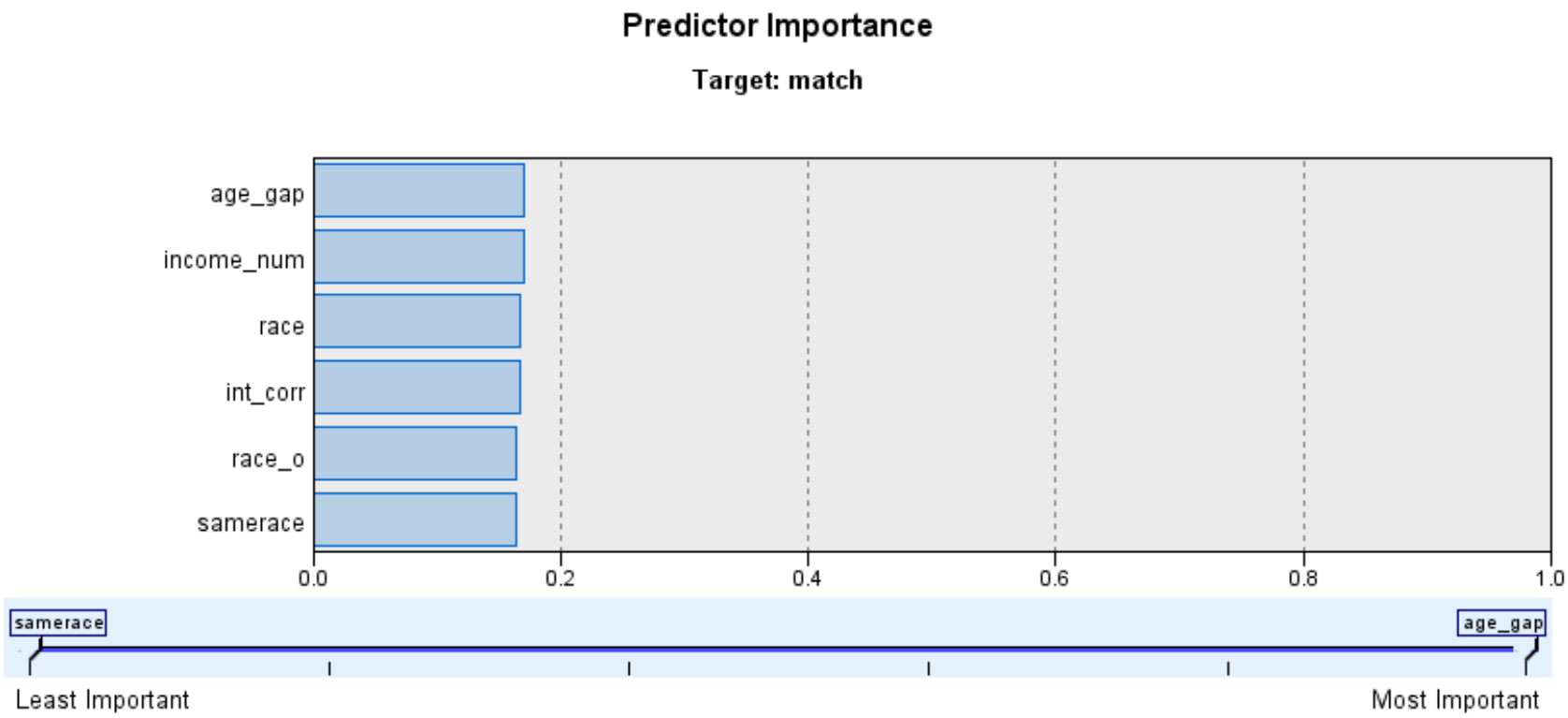
CHAID

Results for output field match

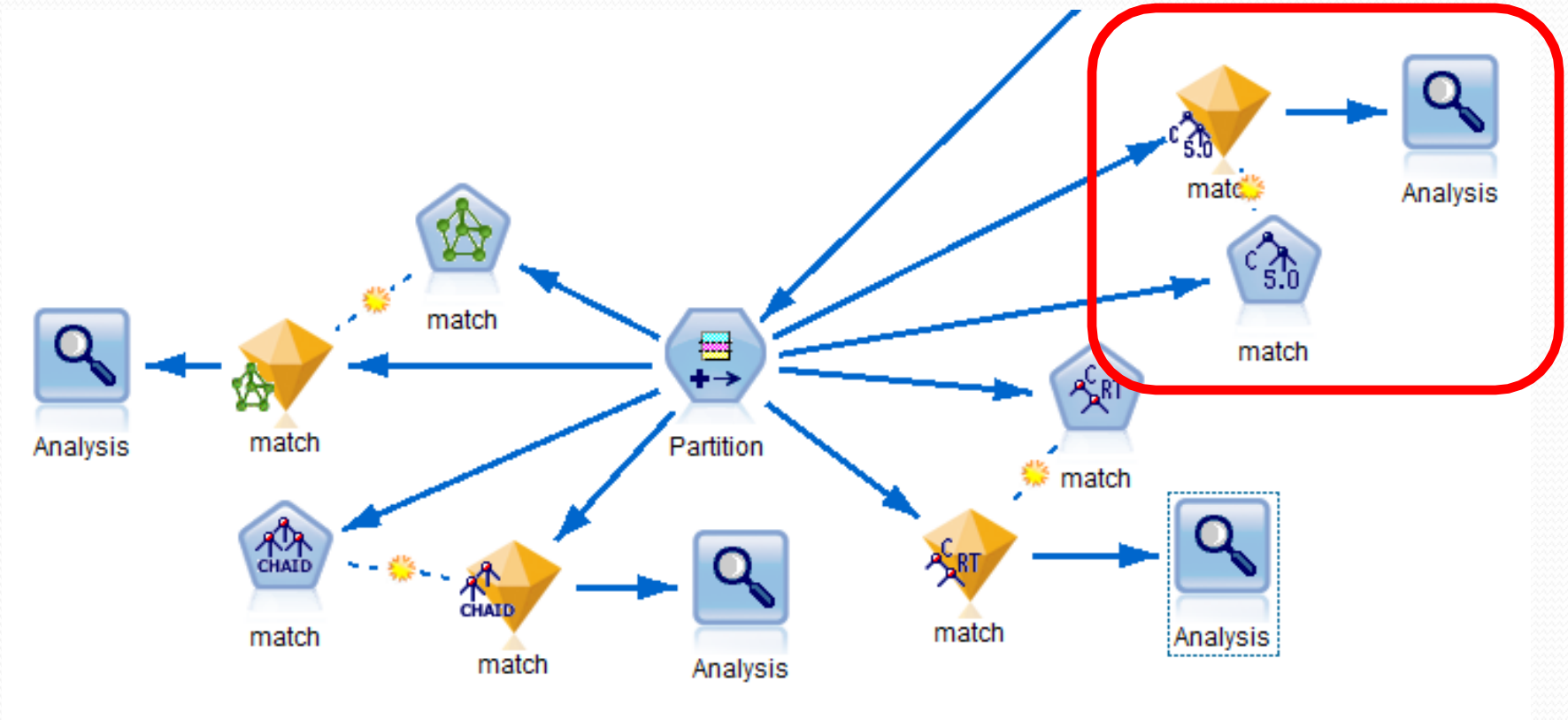
Comparing $R-match with match

'Partition'	1_Training		2_Testing	
Correct	15,114	81.78%	6,536	82.72%
Wrong	3,368	18.22%	1,365	17.28%
Total	18,482		7,901	

CHAID


Predictor Importance
Target: match
age_gap
income_num
race
int_corr
race_o
samerace
0.0
0.2
0.4
0.6
0.8
1.0
samerace
age_gap
Least Important
Most Important

C5.0


Analysis
match
match
Partition
match
Analysis
match
5.0
match
Analysis
match
CHAID
match
CHAID
match
Analysis
match
RT
match
Analysis

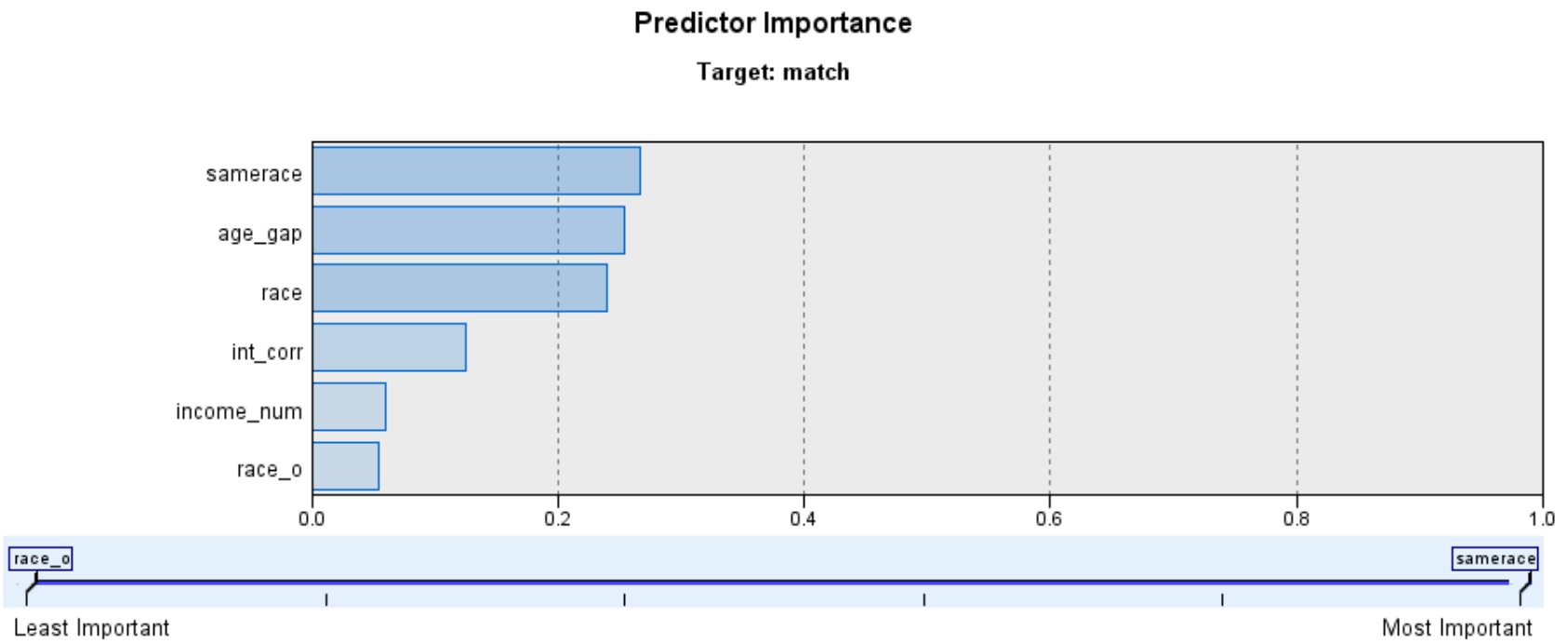
C5.0

Results for output field match

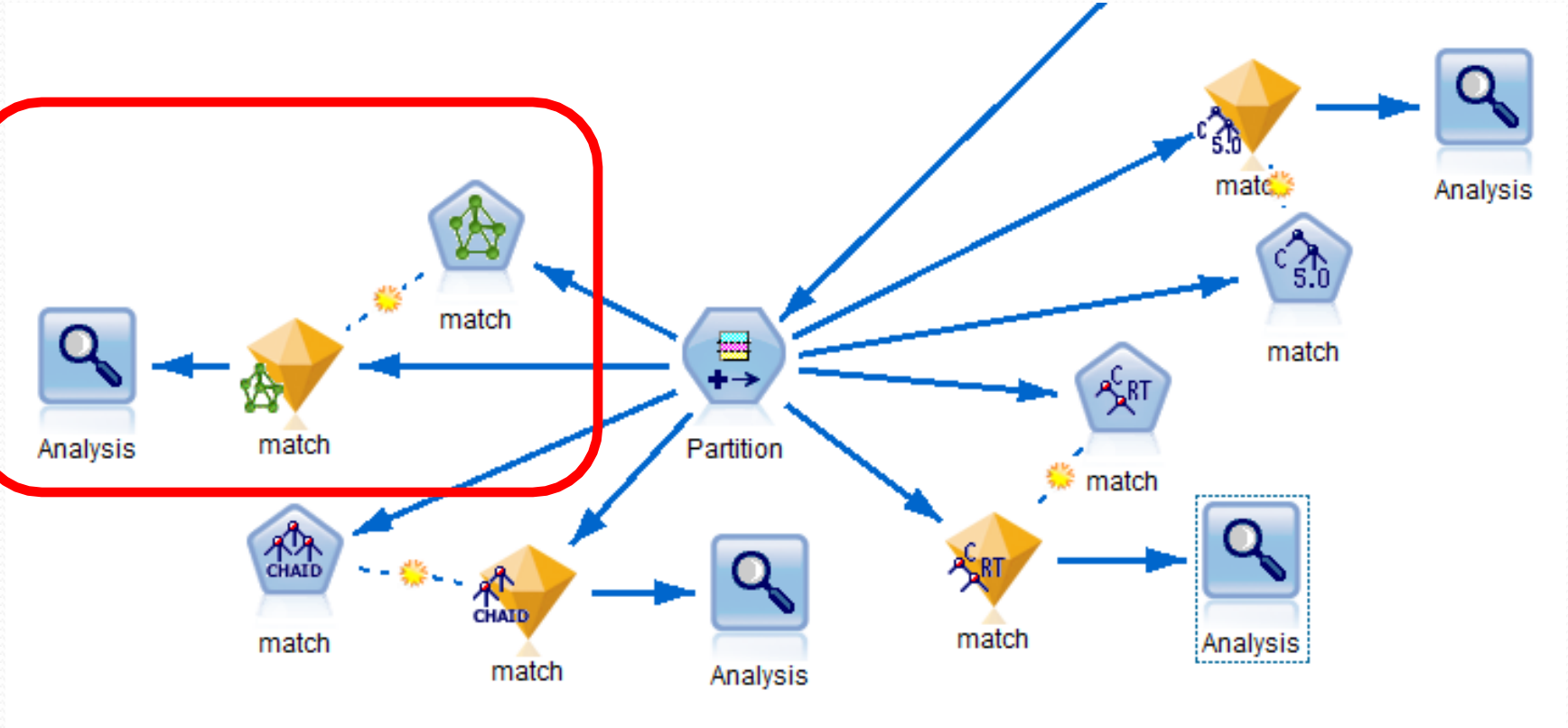
Comparing $C-match with match

'Partition'	1_Training		2_Testing	
Correct	18,573	99.97%	7,933	99.96%
Wrong	6	0.03%	3	0.04%
Total	18,579		7,936	

C5.0


Predictor Importance
Target: match
samerace
age_gap
race
int_corr
income_num
race_o
0.0
0.2
0.4
0.6
0.8
1.0
race_o
samerace
Least Important
Most Important

Neuronska mreža



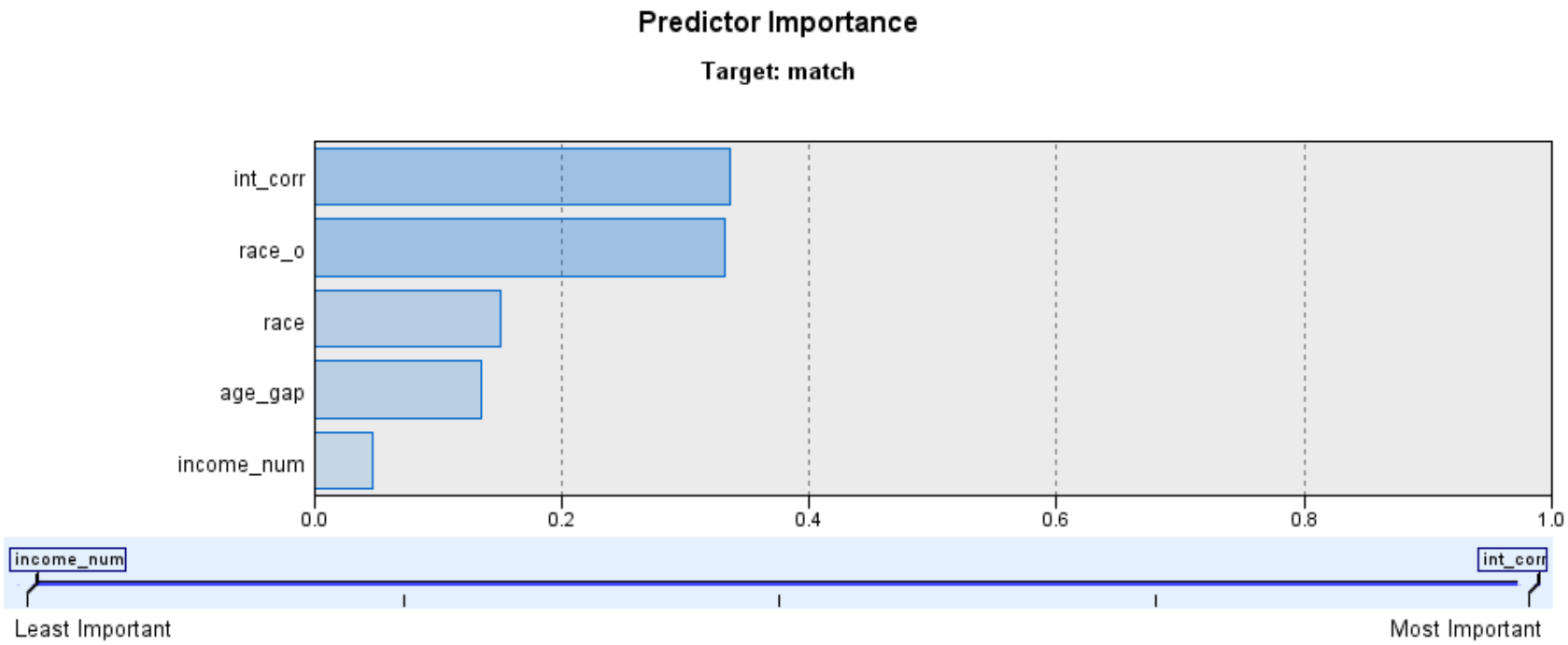
Neuronska mreza

Results for output field match

Comparing $N-match with match

'Partition'	1_Training		2_Testing	
Correct	16,180	87.23%	6,881	86.83%
Wrong	2,369	12.77%	1,044	13.17%
Total	18,549		7,925	

Neuronska mreza



Zaključak

- Najekasnijim se pokazao algoritam C5.0 sa cak 99.96% pogođenih rezultata.
- Ostali algoritmi su se znatno gore pokazali, pri čemu je C&R Tree bio najgori.
- Na osnovu rezultata C5.0 algoritma mozemo zaključiti da su za ispitanike u Kolumbiji od testiranih faktora najbitniji godine i etnička pripadnost, dok su zarada i zajednička interesovanja manje bitni.

Hvala na pažnji!