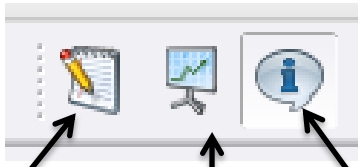
Strojno učenje

Vježbe 1: RapidMiner osnove

Prvo korištenje

- Demo:
  - http://rapid-i.com/content/view/26/84/lang,en/
  - Pokrenite Rapidminer
  - Tri perspektive prikaza:
    - Početna je *Welcome* perspektiva
    - Za početak rada prebacite se na *Design*

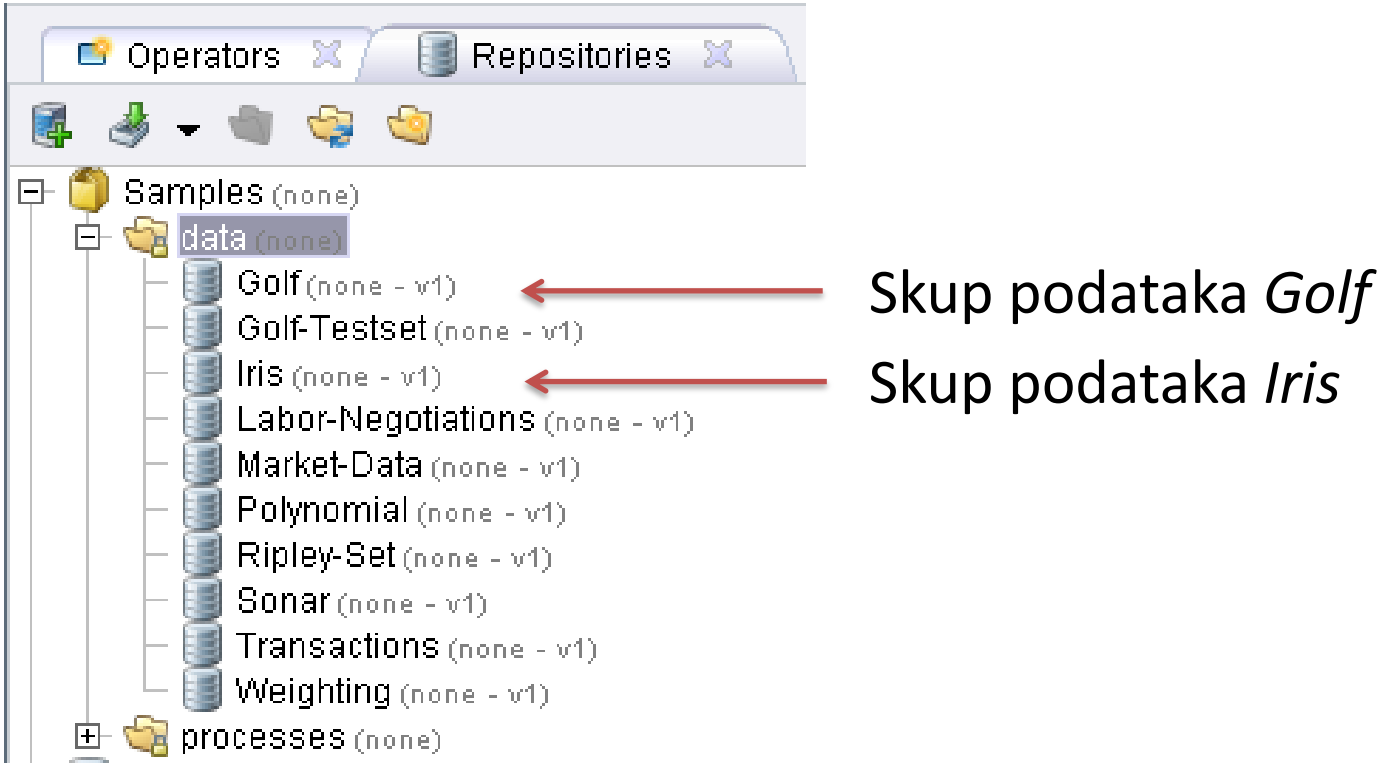


Results *Welcome*

  - Stvorite lokalni repozitorij: VjezbaRM
    - Repozitorij je središnje mjesto za sve vaše **podatke** i procese za **analizu**

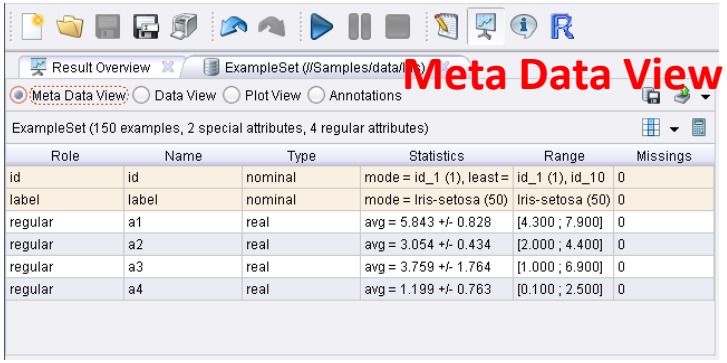
Podaci

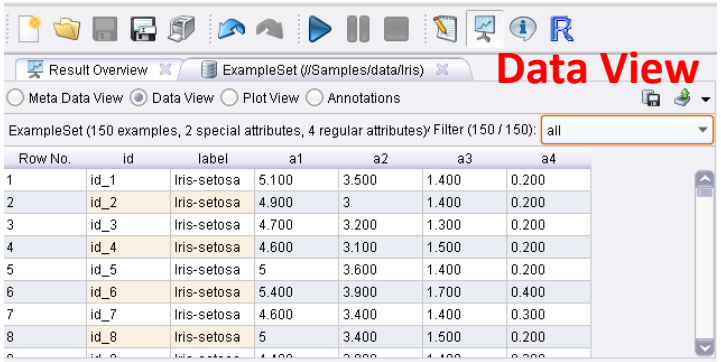
- Rapidminer: Primjeri učitanih podataka
  - Repositories: Samples > data

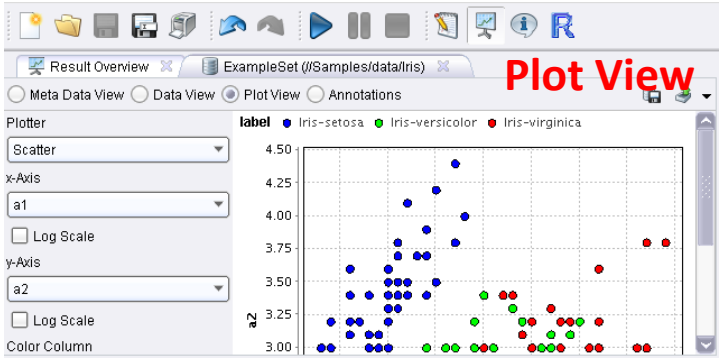


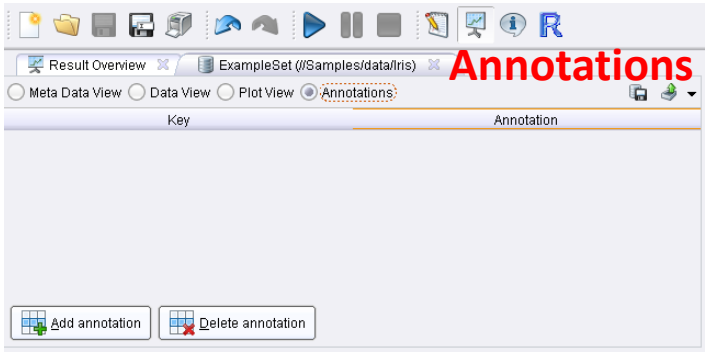
Prikaz podataka

- Dvoklik na učitanu tablicu podataka u repozitoriju prikazuje podatke unutar perspektive *Results*
- Podaci se mogu prikazati na četiri različita načina:









Podaci

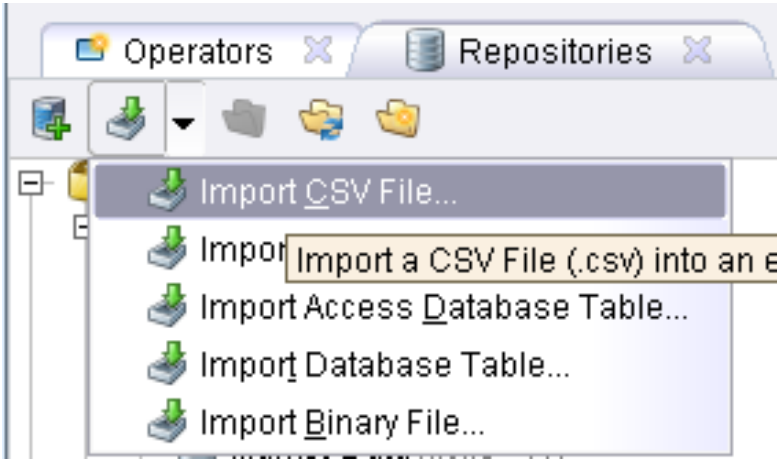
ExampleSet (150 examples, 2 special attributes, 4 regular attributes)

Role	Name	Type	Statistics	Range	Missings
id	id	nominal	mode = id_1 (1	id_1 (1), id_10	0
label	label	nominal	mode = Iris-se	Iris-setosa (50)	0
regular	a1	real	avg = 5.843 +/-	[4.300 ; 7.900]	0
regular	a2	real	avg = 3.054 +/-	[2.000 ; 4.400]	0
regular	a3	real	avg = 3.759 +/-	[1.000 ; 6.900]	0
regular	a4	real	avg = 1.199 +/-	[0.100 ; 2.500]	0

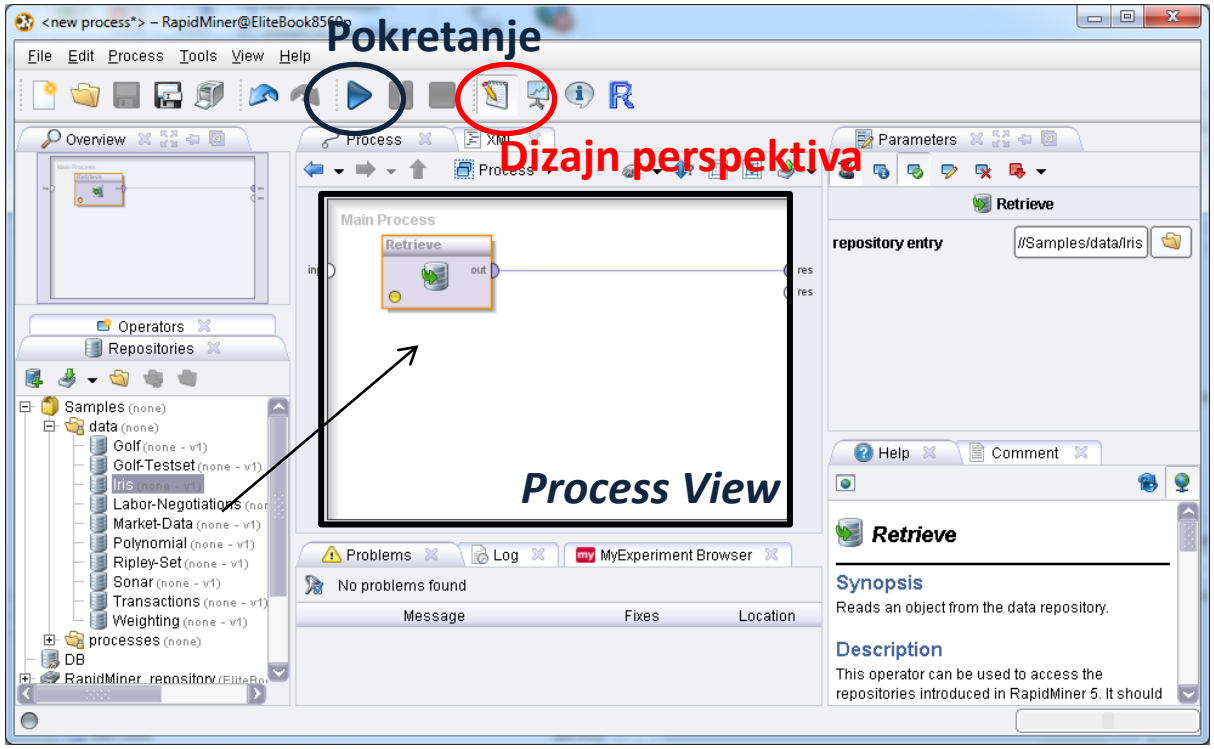
- Uloge atributa (*Role*)
  - **id**: jasno identificira cijeli uzorak
  - **label**: vrijednosti ovog stupca predviđaju se na temlju vrijednosti ostalih stupaca
  - **regular**: atributi koji nemaju neku posebnu ulogu, često ih zovemo: atributi, varijable ,deskriptori, značajke
- Tipovi vrijednosti atributa (*Value Type*)
  - Kategoričke vrijednosti (nominal)
  - Numeričke vrijednosti (numeric):
    - Cijeli brojevi (integer)
    - Realni brojevi (real)
  - Tekst (text)
  - date, date_time, time

Podaci

- **UCI:** http://kdd.ics.uci.edu/
  - http://archive.ics.uci.edu/ml/datasets.html
  - Lenses: http://www.cs.bme.hu/~kiskat/adatb/contact-lenses.arff
  - Učitavanje podataka u *RapidMiner* repozitorij



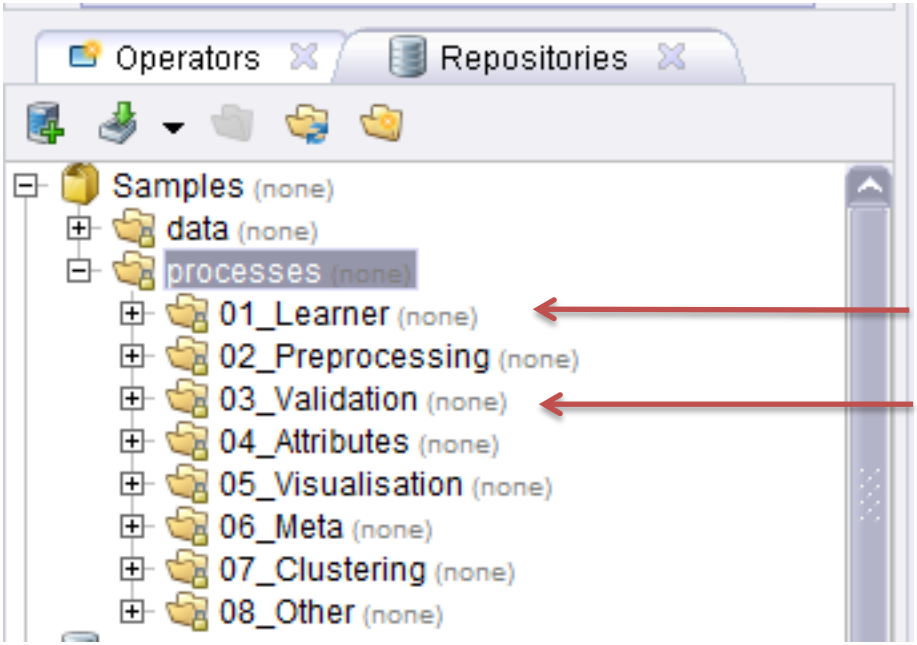
Procesi za analizu



- Odvlačeći skup podataka u prostor za oblikovanje procesa (*Process View*) automatski se stvara ulazni **operator**
- Procesi se definiraju povezivanjem ulaznih/izlaznih operatora
- Konačni rezultat se dovodi do izlaza glavnog procesa

Procesi za analizu

- Rapidminer: Primjeri procesa za analizu
  - Repositories: Samples > processes

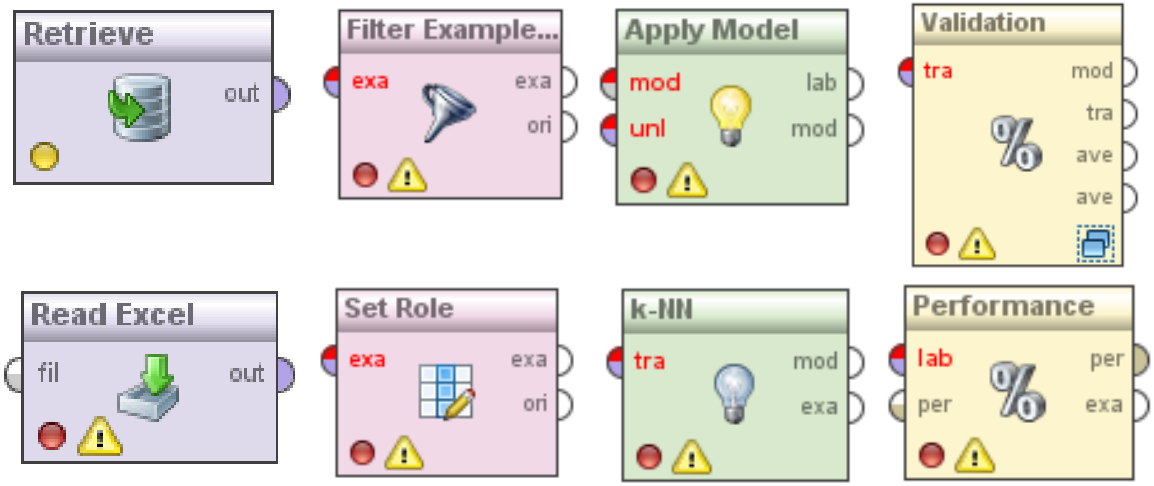


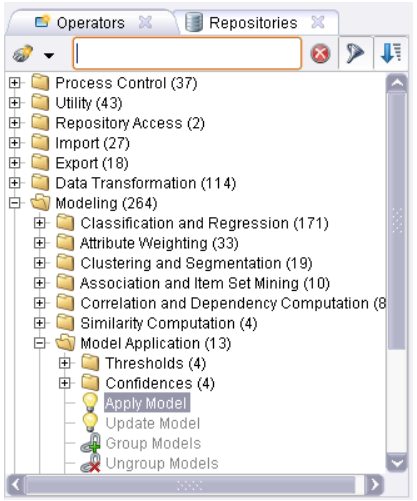
Primjeri za učenje modela

Primjeri za evaluaciju modela

Operatori

- Operatori su osnovni gradivni blokovi RM procesa za analizu
- Svaki operator
  - Ima ulazna i izlazna vrata
  - Definira neku akciju
  - Ovise o postavkama parametara operatora
- Boja operatora određuje njegov tip i boja vrata operatora određuje očekivani ulaz i izlaz
- Operatori različitog tipa se mogu lagano pronaći u prozoru *Operators*





Vježba 1.1: Podaci

- Napravite proces i pohranite ga u svoj repozitorij. Nazovite ga 01_UI_Podaci.
- Preuzmite datoteku contact-lenses.arff
- Iskoristite operator „Read ARFF” za učitavanje skupa podataka iz preuzete datoteke
- Iskoristite operator „Set Role” kako bi postavili atribut „contact-lenses” učitanog skupa podataka na **label**
- Povežite operatore „Read ARFF” i „Set Role” i povežite rezultirani skup podataka s prvim ulazom za rezultate glavnog procesa. Uvjerite se da operatori imaju žutu oznaku (ili zelenu ako je proces već pokrenut).
- Pohranite i pokrenite proces (operator ima dobiva zelenu oznaku)
- Pogledajte prikaz tablice podataka (Data View) učitanog skupa podataka (rezultata pokrenutog procesa)
  - Koliko učitani skup podataka ima primjera?
  - Koliko ima specijalnih, a koliko regularnih atributa?
- Analizirajte prikaz meta podataka učitanog skupa podataka
  - Koliko koje tipovi vrijednosti imaju atributi?
  - Koje osnovne statistike za pojedine atribute možete iščitati iz prikaza?

Vjezba 1.2: Vizualizacija podataka

- Napravite proces i pohranite ga u svoj repozitorij. Nazovite ga 01_UI_PodaciViz.
- Učitajte iz repozitorija primjera RMa skup podataka „Iris”. Povežite njegov izlaz na ulaz rezultata glavnog procesa.
- U prikazu rezultata za vizualizaciju:
  - Nacrtajte „Scatter Plot” u kojem je na apscisi atributi a3, a na ordinati atribut a1, za „Color Column” postavite atribut koji je **label**
  - Vizualno procjenite je li moguće u prostoru koji razapinju ta dva atributa odrediti pravac koji sigurno razdvaja klasu Iris-setosa od ostale dvije klase
  - Nacrtajte „Scatter Plot” a1-a1. Primjetite karakteristiku takvoga grafa.
  - Sada iskoristite „Scatter Matrix” i pogledajte koja dva atributa su najsličnija daju gaf karakteristike najsličnije gornjem grafu? Dakle, korelacija je najveća između ta dva atributa.
  - Iskoristite „3D Scatter Plot Color” i vizualizirajte prostor a1-a2-a3+label, te nakon toga vizualizirajte prostoj a1-a2-a4+label. Jesu li kvalitativno isti?
  - Korištenjem vizualizacije uspjeli samo dobro shvatiti kvalitativnu strukturu podataka. To nam je pomoglo da prostor od 4 atributa svedemo na 3 atributa uz mali gubitak informacija.

Modeliranje – Učenje modela

- RM nam omogućuje da različite modele za učenje koristimo kao crne kutije korištenjem operatora
- Predikcija (nadzirano učenje iz označenih primjera za treniranje):
  - Klasifikacija (učenje grupa)
  - Regresija (učenje funkcija)
- Klasteriranje (nenadzirano grupiranje neoznačenih primjera)

Vježba 2: Stabla odluke

- Otvorite proces 01_UI_Podaci (iz prethdne vježbe) i pohranite kao 02_1_UI_Model_DecisionTree.
- Dodajte operator za učenje „Stablo odlučivanja” (engl. *Decision Tree*)
- Povežite izlaz operatora za učenje s prvim ulazom za rezultate glavnog procesa. Pokrenite proces.
- Analizirajte rezultat procesa
  - Koji atribut se nalazi na krojenu stabla
  - Odredite najpovoljniji tip pomoću dobivenog stabla odluke leće za mladog pacijenta koji normalno stvara suze i koji nema astigmatizam

Korištenje istreniranog modela

- Dosada smo vidjeli tek kako se modeli generiraju iz danih podataka
- Vidjeli samo da sami opis modela može biti koristan za bolje uočavanje strukture podataka
- No, sada ćemo primijeniti naučene modele na „novim" podacima → taj proces nazivamo prognozom ili predikcijom
- Jedan od najvažnijih operatora u RMu je operator „**Apply Model**"
- Taj operator uzima ulazni <Skup podataka> i <Model> i primjenjuje model na taj skup podataka
- Rezultat će biti dani skup podataka, no sada će se najmanje jedan novi stupac dodati u tablicu tog skupa podataka
- Taj novi stupac sadrži određene klase (ili numeričke vrijednosti u slučaju regresije) i zato se zove prognoza(engl. *prediction*)

Primjena modela

- Dodatno uz predikciju, **klasifikacijski modeli** će također generirati stupce pouzdanosti (engl. confidence)
- Pouzdanost kvantificira vjerojatnost da, s obzirom na model, određena vrijednost bude predviđena za dani primjer
- Vrijednosti pouzdanosti određene su za sve moguće **klase**
- Pouzdanosti se sumiraju u 1
- Pouzdanosti nisu nužno jednake vjerojatnosti te vrijednosti za dani primjer

Vježba 3: Primjena modela

- Stvorite proces i nazovite ga 03_1_UI_Model_DecisionTree.
- Preuzmite skup podataka Iris iz repozitorija primjera
- Dodajte operator za učenje „Stablo odlučivanja” (engl. *Decision Tree*)
- Dodajte operator „Apply Model” i povežite izlaze operatora za učenje s operatorom za primjenjivanje modela
- Analizirajte rezultat procesa
  - Koji su stupci stvoreni? Za što služe?
- Zamijenite model učenja u Naive Bayes i pospremite proces kao 03_02_UI_Model_NaiveByes.
  - Kako su se rezultati predikcije promijenili?

Evaluacija modela

- Kako mjerimo kvalitetu modela - mjera za evaluaciju modela?
- Funkcija za izračunavanje greške koju model radi na skupu podataka za testiranje modela se zove mjera za evaluaciju modela (engl. *performanse criterion*)
- Skup mjera za evaluaciju modela u RapidMineru se naziva vektor performanse (engl. performance vector)

Evaluacija modela za klasifikaciju

- Zadatak evaluacije klasifikacijskih modela je izmjeriti u kojem stupnju klasifikacija sugerirana izrađenim modelom odgovara stvarnoj klasifikaciji
- Ovisno o načinu promatranja performansi modela postoji više različitih mjera za evaluaciju modela
- Mjera za evaluaciju klasifikacijskih modela u RMu:
  - Točnost (engl. *Accuracy*)
  - Greška (engl. Error)
  - Kapa statistike
  - Spearman Rho
  - Kendall Tau
  - *Root Mean Squared Error*
- Za binarnu klasifikaciju:
  - Preciznost (engl. *Precision*)
  - Odziv (engl. *Recall*)
  - ROC, AUC (Area Under Curve)

Evaluacija modela za klasifikaciju

- Greška = pogrešna klasifikacija primjera
- Evaluacija klasifikatora: koliko je klasifikator za promatrani skup primjera napravio grešaka?

$$Točnost = \frac{broj\ ispravno\ klasificiranih\ primjera}{ukupan\ broj\ primjera}$$

- Dva su osnovna nedostatka *točnosti* kao mjere za evaluaciju
  - Zanemaruju se razlike između tipova grešaka
  - Zavisna je o distribuciji klasa u skupu podataka, a ne o karakteristikama primjera
- U većini primjena je važno razlikovati tipove greške: npr. sustav za detekciju neželjene pošte (Spam) može pogrešno označiti željenu poruku (Ham) kao Spam, a Spam kao Ham, u tome slučaju greška označavanja Spama kao Hama ima puno manju težinu nego greška označavanja Hama kao Spama
- Razlikovanje više tipova greške:
  - rezultat klasifikacije => MATRICE GREŠKE

Evaluacija modela za klasifikaciju

Matrica grešaka za klasifikacijski problem s dvije klase

Contigency Table; Confusion Matrix	Pozitivni primjeri klase C	Negativni primjeri klase C
Pozitivna prognoza klase C	Stvarno pozitivni (TP)	Lažno pozitivni (FP)
Negativna prognoza klase C	Lažno negativni (FN)	Stvarno negativni (TN)

Četiri različita rezultata prognoze: TP, FP, FN, TN

$$Točnost = \frac{broj\ ispravno\ klasificiranih\ primjera}{ukupan\ broj\ primjera} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Greška = \frac{?}{?}$$

$$Senzitivnost = \frac{TP}{TP + FN}$$ → točnost u pozitivnim primjerima

$$Specifičnost = \frac{TN}{TN + FP}$$ → točnost u negativnim primjerima

Evaluacija modela za klasifikaciju

Matrica grešaka za klasifikacijski problem s dvije klase

Contigency Table; Confusion Matrix	Pozitivni primjeri klase C	Negativni primjeri klase C
Pozitivna prognoza klase C	Stvarno pozitivni (TP)	Lažno pozitivni (FP)
Negativna prognoza klase C	Lažno negativni (FN)	Stvarno negativni (TN)

Četiri različita rezultata prognoze: TP, FP, FN, TN

Ako je TP << TN, onda češće koristimo mjere za evaluaciju modela:

$$Odziv = \frac{TP}{TP + FN}$$ → točnost u pozitivnim primjerima (jednako kao Senzitivnost)

$$Preciznost = \frac{TP}{TP + FP}$$ → točnost u pozitivnoj prognozi ciljne klase

Evaluacija modela za klasifikaciju

Matrica grešaka za klasifikacijski problem s dvije klase

Contigency Table; Confusion Matrix	Pozitivni primjeri klase C	Negativni primjeri klase C
Pozitivna prognoza klase C	Stvarno pozitivni (TP)	Lažno pozitivni (FP)
Negativna prognoza klase C	Lažno negativni (FN)	Stvarno negativni (TN)

Četiri različita rezultata prognoze: TP, FP, FN, TN

Mjere za evaluaciju klasifikacijskih modela zasnovanih na principima preciznosti:

$$Pozitivna\ prediktivna\ vrijednost = \frac{TP}{TP + FP}$$ → Preciznost u pozitivnim primjerima

$$Negativna\ prediktivna\ vrijednost = \frac{TN}{TN + FN}$$ → Preciznost u negativnim primjerima

Evaluacija modela za klasifikaciju

Matrica grešaka za klasifikacijski problem s dvije klase

Contigency Table; Confusion Matrix	Pozitivni primjeri klase C	Negativni primjeri klase C
Pozitivna prognoza klase C	Stvarno pozitivni (TP)	Lažno pozitivni (FP)
Negativna prognoza klase C	Lažno negativni (FN)	Stvarno negativni (TN)

Četiri različita rezultata prognoze: TP, FP, FN, TN

- Prije navedeni parovi mjera evaluacije modela (Senzitivnost/Specifičnost;…) pokazuju specifičnu točnost klasifikacijskog modela s komplementarnih pogleda
- Često kvalitetu želimo izraziti jednim brojem => fiksiramo vrijednost jedne mjere i uz taj uvjet promatra se vrijednost samo druge mjere ili češće usrednjavanjem jedne od mjera po više fiksiranih vrijednosti druge mjere (npr. preciznost uz fiksiranu vrijednost odaziva od 20%, 50% i 80% - srednja preciznost u 3 točke)
- Primjer mjere koja se ne zasniva na fiksiranju jedne komponente para mjera:

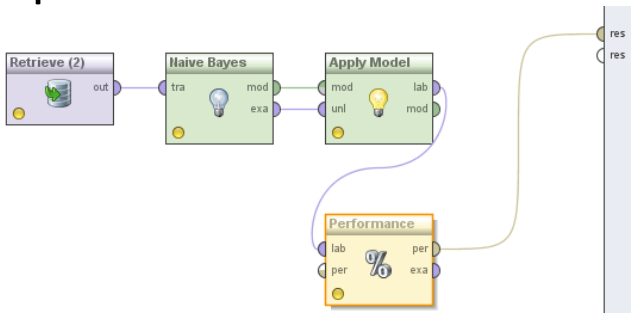
$$F - mjera = \frac{2 \times odaziv \times preciznost}{odaziv + preciznost} = \frac{2TP}{2TP + FP + FN}$$

ROC analiza

- Želimo opisati obuhvatnije opisati međusobne ovisnosti različitih mjera i/ili parametara
- ROC = *Receiver Operating Characteristic*
  - Obuhvatna, grafički orijentirana mjera kvalitete klasifikacijskog modela

Vježba 4.1. Evaluacija modela klasifikacije

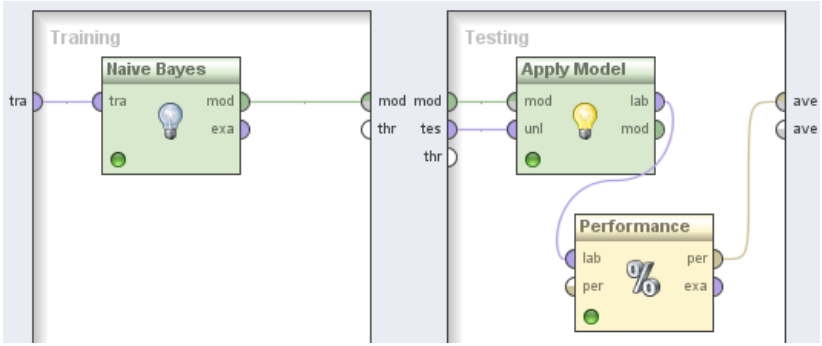
- Napravite novi proces i nazovite ga 04_UI_Evaluacija_TestTrain
- Iz repozitorija primjera RM uzmite skup podatka „Golf”, operatore „Naive Bayes”, „Apply Model” i „Performanse (Classification)” i spojite ih u proces prikazan na slici



- Pokretanjem procesa dobili smo koliko dobro istrenirani model klasificira skup podatka na kojima je učio.
- Uzmite iz repozitorija primjera RM skup podataka „Golf testset” i spojite ga na ulaz unl operatora Apply Model (umjesto izlaza exa operatora Naive Bayes)
- Pokretanjem procesa dobili smo koliko dobro istrenirani model klasificira neviđeni skup podatka odnosno podatke na kojima nije učio. Uočite razliku u greškama.

Vježba 4.2: Evaluacija modela klasifikacije

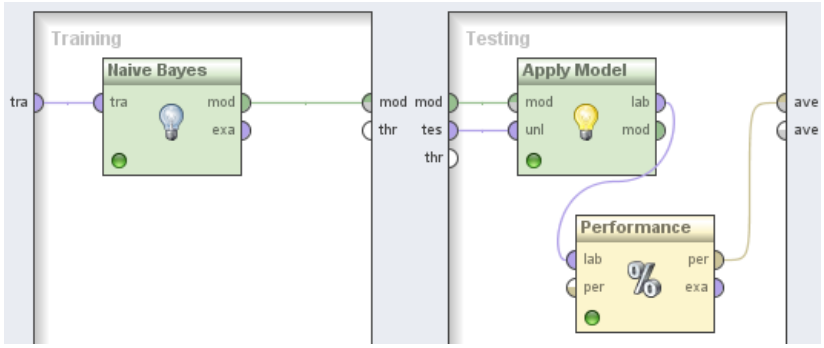
- Napravite novi proces i nazovite ga 04_UI_Evaluacija_SplitValidation
- Iz repozitorija primjera RM uzmite skup podatka „Labor-Negotiations” i operator „Split Validation”
- „Split Validation” je operator koji ugnježđuje druge operatore podijeljen u dva procesa, realizirajte ga kao na slici:



- Validacijom modela procijenili smo grešku klasifikacije na neviđenim podacima (kvalitetu generalizacije)
- Primijetite rezultat operatora SplitValidation

Vježba 4.3: Evaluacija modela klasifikacije

- Napravite novi proces i nazovite ga 04_UI_Evaluacija_XValidation
- Iz repozitorija primjera RM uzmite skup podatka „Labor-Negotiations” i operator „XValidation”
- „XValidation” je operator koji ugnježđuje druge operatore podijeljen u dva procesa, realizirajte ga kao na slici:



- Validacijom modela procijenili smo grešku klasifikacije na neviđenim podacima (kvalitetu generalizacije)
- Usporedite rezultat operatora XValidation s rezultatima SplitValidation operatora iz prošlog primjera
- Što mislite, koja je procjena bolja?