Istraživanje podataka 1 - SPSS

SPSS (Statistical Package for the Social Sciences) koristi se za učitavanje i manipulaciju podacima, kao i izvoz rezultata. Operacije koje se mogu primeniti nad podacima su predstavljene kao čvorovi, a niz povezanih operacija (čvorova) naziva se **tok podataka (data stream)**. Vezama između čvorova određuje se pravac toka podataka.

Za učitavanje podataka iz csv fajla koristi se čvor **Var File**. Opcije:

- **File** - učitavanje fajla i opcije.
- **Data** - prikaz atributa i njihovi tipovi.
- **Filter** - odabir atributa.
- **Types** - detaljne informacije o atributima:
  1. **Measurement** - tip upotrebe atributa:
     - **Default** - nepoznat
     - **Continuous** - neprekidan
     - **Categorical** - kategorički, nakon čitanja vrednosti prelazi u Flag, Nominal ili Typeless
     - **Flag** - binarni
     - **Nominal** - imenski
     - **Ordinal** - redni
     - **Typeless** - atributi koji imaju jednu vrednost ili imenski atributi sa više vrednosti od dozvoljenog
  2. **Values** - interval ili moguće vrednosti atributa:
     - **Read** - informacije se učitavaju pri izvršavanju čvora.
     - **Read+** - informacije se učitavaju i dodaju već definisanim ako postoje.
     - **Pass** - ne učitavaju se informacije.
     - **Current** - ostaju već definisane vrednosti.
     - **Specify** - posebno definisanje vrednosti.
  3. **Missing** - definisanje načina obrade nedostajućih vrednosti.
  4. **Check** - definisanje akcije za objekte koji imaju vrednost koja ne pripada definisanom intervalu ili listi vrednosti u Values.
     - **None** - ne menja se vrednost.
     - **Nullify** - postavlja se na null.
     - **Coerce** - vrednost se prebacuje u legalnu. Za flag u false, za nominal i ordinal u prvu vrednost iz skupa, za neprekidne u gornju/donju granicu ako je vrednost veća/manja od gornje/donje granice ili u srednju vrednost ako je null.
     - **Discard** - ceo slog se odbacuje.

    - **Warn** - prijavljuje se broj slogova sa nepravilnim vrednostima.
    - **Abort** - prijavljuje se greška.
5. **Role** - da li je ciljna (**Target**) ili ulazna (**Input**) promenljiva.

Za učitavanje xlsx fajlova koristi se čvor **Excel**.

Table je čvor koji omogućava tabelarni prikaz podataka.

Select je čvor preko koga izdvajamo podatke. Izaberemo atribut(e) i zadajemo neke uslove koje moraju da ispune ili koje ne smeju da ispune.

Data Audit je čvor za upoznavanje sa podacima. Prikazuje sumarne statistike za atribute i grafike sa distribucijom vrednosti po atributima. Prikazuje i izveštaj o nedostajućim vrednostima, elementima van granica, ekstremnim vrednostima i omogućava definisanje akcija za obradu tih vrednosti. Opcije:

- **Settings** - biranje statistika.
- **Quality** - podešavanje načina na koji se računaju autlajeri i ekstremne vrednosti. Može biti udaljenost od očekivanja u jedinici standardne devijacije ili udaljenost preko kvartila.

Nedostajuće vrednosti su Null ili sistemski nedostajuće vrednosti. Označene su sa $null$. Prazne niske i beline su regularne vrednosti i mogu se definisati kao **blanko vrednosti**. Moguće je i prisustvo korisnički definisanih nedostajućih vrednosti u nekim čvorovima. Koristi se čvor **Filler**. Metode obrade nedostajućih vrednosti mogu biti:

- **Fixed** - zamena vrednošću koja je zadata konstanta ili rezultat izabrane statistike.
- **Random** - zamena izborom slučajne vrednosti.
- **Expression** - zamena rezultatom izraza.
- **Algorithm** - zamena korišćenjem vrednosti predviđene modelom dobijenog algoritmom C&RT.
- **Set Globals** - računa statistike atributa koje se mogu kasnije koristiti globalno u drugim čvorovima.
- **Type** - prikaz informacija o tipovima atributa.
- **Binning** - diskretizacija.
- **SetToFlag** - binarizacija. Označimo **aggregate keys** da bi se spojile stavke po transakcijama.
- **Reclassify** - spajanje nekoliko različitih vrednosti atributa u jednu novu ili već postojeću vrednost.

Stabla odlučivanja:

- **C5.0** - klasifikacija sa korišćenjem informacione dobiti i entropije. Za numeričke atribute koristi binarnu podelu, a za kategoričke po jednu granu za svaku moguću vrednost (moguće je i grupisanje).
- **Partition** - podela podataka na trening i test skup. Opcije:

    - **Boosting** - puta pravi model, gde svaki naredni pokušava da ispravlja greške prethodnog.
    - **Cross-validation** - **unakrsna validacija**, tj. podaci se dele na grupa i model se puta trenira pri čemu svaki put sledeća grupa predstavlja test skup, a preostali podaci trening skup.
- **Analysis** - ispisuje informacije o modelu. Moguće je uključiti i **matrice konfuzije** koje prikazuju koliko instanci koje klase je tačno klasifikovano, a koliko nije i u koju klasu su pogrešno raspoređene.

KNN: Modeling Supervised KNN

Pravila pridruživanja: Modeling Association Apriori:

- Use custom field assignments transaction format (ID transakcija, content proizvodi)
- Use predefined roles (ako smo uradili binarizaciju i označili id transakcije kao tip None, a tipove za stavke kao Both (mogu biti i glava i rep pravila)). Možemo povezati i **Web** i označiti show true flags only za prikaz povezanosti stavki.
- Preko "filter" dugmeta izvlačimo samo pravila koja su vezana za neke atribute. Klikom na generate možemo da sačuvamo poseban model koji sadrži samo ta izabrana pravila.