Úloha zpracování a analýzy dat

Zadání

Na zadaných datech proveďte analýzu, která bude obsahovat roztřídění dat do skupin v závislosti na typu podávaného léku, na výsledku léčby primární choroby (primární pro stanovení účinnosti léku), přítomnosti jiné medikace, věku, průměrného krevního tlaku [1] a BMI [2] (nazvěme tyto 4 proměnné faktory) a bude zohledňovat výskyt sekundárních chorob před a po léčbě. Analýza bude obsahovat sledované pacienty, kteří se jedním či druhým lékem vyléčili, tak ale i ty, kteří se nevyléčili. Zároveň by analýza měla ve všech skupinách (rozdělení podle faktorů) i pro výsledný průnik obsahovat statistické testování pro dokázání (vyvrácení), zda je některý z léků účinnější, nebo zda mohly rozdíly mezi léky vzniknout náhodou (vždy na hladině významnosti 0.05) [3]. Diskutujte případný vliv věku, průměrného tlaku, BMI či přítomnosti jiné medikace na výsledek léčby (ve všech variantách). Zároveň se zaměřte na nalezení zajímavých faktů v datech a jejich potvrzení (např. že pacienti v rozpětí jistého BMI mají podobný tlak). Analýzu je doporučené zpracovávat v MATLABu, při domluvě lze však využít i jiných nástrojů. Do závěrečného shrnutí zařaďte rozdělení do jednotlivých skupin pomocí souboru pravidel nebo rozhodovacího stromu.

Data

Data jsou v souboru formátu csv v kódování ASCII. Data obsahují 11 atributů a 10000 instancí. Atributy jsou následující:

1. věk [roky]
2. průměrný krevní tlak [mmHg]
3. BMI [-]
4. výskyt primární choroby před léčbou [0,1]
5. výskyt první sekundární choroby před léčbou [0,1]
6. výskyt druhé sekundární choroby před léčbou [0,1]
7. přítomnost jiné medikace [0,1]
8. použití léku typu 1 nebo 2 [1,2]
9. výskyt primární choroby po léčbě [0,1]
10. výskyt první sekundární choroby po léčbě [0,1]
11. výskyt druhé sekundární choroby po léčbě [0,1]

Atributy 1, 2 a 3 jsou přirozená čísla. Atributy 4, 5, 6, 7, 9, 10 a 11 nabývají hodnot 0 (není přítomna nemoc či jiná mediace) nebo 1 (je přítomna nemoc či jiná medikace). Atribut 8 nabývá hodnot 1 (první lék) nebo 2 (druhý lék). Každý pacient dostával právě jeden vybraný lék, tedy žádný pacient nebyl léčen zároveň oběma léky. O závažnosti jednotlivých chorob nevíme nic. O vazbě chorob na věk, krevní tlak či BMI také nejsou žádné informace (nemůžeme tuto vazbu předpokládat). U jiné medikace také nemáme informace o počtu léků a jejich účelu.

Co tedy má analýza obsahovat a co se hodnotí

- Popisná analýza dat jako celku a v závislosti na rozdělení věku, průměrného krevního tlaku, BMI a výskytu jiné medikace: tj. četnostní analýza a popis, souhrn výsledků u primární a sekundárních chorob a nalezení zajímavých faktů v datech;
  hodnoceno 0 – 6 body.
- Shrnutí analýz a nalezení průniků mezi jednotlivými skupinami a vytvoření rozhodovacích pravidel (stromu) pro jednotlivé léky;
  hodnoceno 0 – 6 body.
- Statistické vyhodnocení účinnosti léků v rámci jednotlivých skupin a/i pro celkové zhodnocení (průnik jednotlivých skupin) tvořící rozhodovací pravidla, Statistické zhodnocení faktů nalezených v datech;
  hodnoceno 0 – 6 body.
- Zpracování závěrečné zprávy: úroveň zpracování textové části, úroveň zpracování grafické části, přehlednost, jasnost, správnost;
  hodnoceno 0 – 4 body.

Hodnocena je primárně textová zpráva popisující vaši práci a obsahující vaše poznatky, zhodnocení a závěry. Zdrojový kód slouží k možnosti dopátrat se, na základě jakého procesu jste dospěli k vámi prezentovaným výsledkům a k ověření pravosti a originality vaší práce.

Odevzdání

Úlohu odevzdáte přes **UploadSystem** systému CourseWare do **8. 11. 2015**.
Odevzdávat budete zip archiv, který bude obsahovat:

- zdrojový kód (kód zpracování dat)
- zprávu, která bude obsahovat (v textové a grafické podobě) souhrn vašeho postupu a všechny vaše poznatky a závěry.

Doporučené zdroje

[1] http://en.wikipedia.org/wiki/Blood_pressure
[2] http://en.wikipedia.org/wiki/Body_mass_index
[3] Jana Zvárová: Základy statistiky pro biomedicínské obory.