Úloha zpracování 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) 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). 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 (např. Skupina 1: věk = 30 – 50 & průměr krevního tlaku = 100 – 120 & BMI = 20 – 24 & prim. choroba = 1 & 1. sek. choroba = 0 & 2. sek. choroba = 0 & jiná medikace = 0 & lék = 1 & prim. choroba = 0 & 1. sek. choroba = 1 & 2. sek. choroba = 0).

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 12 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.

Odevzdání

Úlohu odevzdáte přes UploadSystem systému CourseWare do 20. 3. 2012. Odevzdávat budete zdrojový kód (kód zpracování dat) a zprávu, která bude obsahovat (v textové a grafické podobě) souhrn vašeho postup a všechny vaše poznatky a závěry.

Hodnocení

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

Minimální obsah

- Analýza dat jako celku: četnostní analýza (analýza kategoriálních dat), souhrn výsledků u primární a sekundárních chorob, statistické vyhodnocení účinnosti léků; hodnoceno 0 – 3 body.

- Analýza dat v závislosti na rozdělení věku: nalezení skupin dle daného faktoru, četnostní analýza v rámci skupin, souhrn výsledků u primární a sekundárních chorob v rámci skupin, statistické vyhodnocení účinnosti léků v rámci skupin; hodnoceno 0 – 3 body.

- Analýza dat v závislosti na průměrném krevním tlaku: nalezení skupin dle daného faktoru, četnostní analýza v rámci skupin, souhrn výsledků u primární a sekundárních chorob v rámci skupin, statistické vyhodnocení účinnosti léků v rámci skupin; hodnoceno 0 – 3 body.

- Analýza dat v závislosti na BMI: nalezení skupin, četnostní analýza v rámci skupin dle daného faktoru, souhrn výsledků u primární a sekundárních chorob v rámci skupin, statistické vyhodnocení účinnosti léků v rámci skupin; hodnoceno 0 – 3 body.

- Analýza dat v závislosti na výskytu jiné medikace: nalezení skupin dle daného faktoru, četnostní analýza v rámci skupin, souhrn výsledků u primární a sekundárních chorob v rámci skupin, statistické vyhodnocení účinnosti léků v rámci skupin; hodnoceno 0 – 3 body.

- Shrnutí analýz a vytvoření rozhodovacího stromu, resp. souboru pravidel pro stanovení rozhodnutí o léčbě lékem 1 nebo lékem 2: Souhrn analýz ze všech rozdělení, nalezení průniků a testování účinnosti na těchto množinách, zohlednění sekundárních chorob, sestavení rozhodovacího algoritmu; hodnoceno 0 – 4 body.

- Zpracování závěrečné zprávy: úroveň zpracování textové části, úroveň zpracování grafické části, přehlednost, jasnost, správnost; hodnoceno 0 – 3 body.

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, http://ucebnice.euromise.cz/index.php?conn=0&section=biostat1