Zadání semestrální práce
Téma: cukrovka

Cíl

Cílem této práce je analyzovat data nasbíraná na vzorku populace Indiánek kmene Pima, na kterém byla zkoumána cukrovka. Vaším úkolem bude:

- odlišit ženy trpící cukrovkou od žen zdravých,
- modelovat krevní tlak v závislosti na hodnotě *body mass index (BMI)*,
- pokusit se nalézt vnitřní strukturu v datech.

Data

Data byla nasbírána po roce 1965 na vzorku několika set Indiánek kmene Pima poblíž Phoenixu v Arizoně. U žen byl zjištěn věk, rodinná zátěž (predispozice k cukrovce), body mass index (BMI) a další příznaky (viz Tabulka 1), a byla u nich zjištěna přítomnost cukrovky.

Data jsou k dispozici v souboru `diabetes.csv`.

`n.preg`	počet těhotenství
`glu`	hladina glukózy v krvi (v rámci orálního testu na glukózovou toleranci)
`bp`	diastolický krevní tlak [mm Hg]
`fold`	tloušťka kožní řasy na tricepsu [mm]
`bmi`	*body mass index* (hmotnost [kg]/(výška [m])^2)
`ped`	rodinná zátěž (predispozice k cukrovce zjištěná z příbuzenstva)
`age`	věk [roky]
`diab`	příznak, zda u ženy byla diagnostikována cukrovka

Tabulka 1: Popis naměřených příznaků.

Požadované kroky analýzy

1. Průzkumná analýza. **[5 bodů]**
   - Kolik máte k dispozici dat (kolik osob, kolik příznaků)?
   - Obsahují data nějaké chybějící hodnoty? Pokud ano, jak se s nimi vypořádáte?
   - Jsou v datech nějaká odlehlá pozorování? Pokud ano, jak se s nimi vypořádáte?
   - Je třeba data pro účely následujícího zpracování transformovat? Pokud ano, jak?
   - S ohledem na následující úkoly vizualizujte vybrané příznaky resp. vztahy mezi příznaky resp. vztahy mezi příznaky a výskytem cukrovky.
2. Klasifikace žen na nemocné cukrovkou a zdravé. **[10 bodů]**
   - Vyberte alespoň dva klasifikátory vhodné pro tento typ dat. Podle čeho budete vybírat?
   - Na vhodné podmnožině dat vybrané klasifikátory natrénujte. Jaké úspěšnosti dosahují? Podle jakých příznaků se rozhodují? Dává to smysl?

- Jakou úspěšnost klasifikace očekáváte v hypotetickém reálném nasazení Vašeho klasifikátoru, tj. v případě klasifikace nově příchozí ženy?
- Jaká je pravděpodobnost, že nově příchozí žena bude trpět cukrovkou?
- Jaká je pravděpodobnost, že nově příchozí nemocná žena bude klasifikována jako zdravá? Jaká bude naproti tomu pravděpodobnost, že nově příchozí nemocná žena bude klasifikován jako zdravá? Kterou chybu považujete za závažnější a proč?
- Má ve vašem případě na přesnost klasifikace vliv to, zda je trénovací (testovací) množina vyvážená? Pokud ano, jaký? Vyvažovali jste trénovací (testovací) množinu? Pokud ano, proč a jak? Pokud ne, proč ne?

3. Modelování závislosti krevního tlaku na BMI. [**3 body**]

   - Vhodnou metodou danou závislost modelujte a formálně ji zapište. Je závislost statisticky významná? Výsledek interpretujte.
   - Existuje vztah hodnoty krevního tlaku a jiných příznaků? Jaká kombinace příznaků podle Vás nejlépe modeluje hodnotu krevního tlaku? Je vliv BMI na hodnotu krevního tlaku v takovém modelu jiný, než v modelu uvažujícím izolovaný vliv BMI?

4. Hledání podskupin v datech. [**4 body**]

   - Pokuste se v datech *pouze na základě hodnot příznaků* (tj. bez zohlednění toho, zda se jedná o zdravou či nemocnou ženu) nalézt vnitřní strukturu (podskupiny). Zamyslete se, zda je třeba data normalizovat. Výsledky vizualizujte a interpretujte. Lze popsat nalezené shluky např. typickými/průměrnými reprezentanty?
   - Je možno dát nalezenou strukturu do souvislosti s rozlišením žen na nemocné a zdravé? (Jinými slovy: neodpovídají nalezené podskupiny do značné míry nemocným/zdravým ženám?) Vizualizujte a/nebo popište tabulkou.

Zpráva o řešení

Svoji práci shrňte ve formě sdělení ve formátu PDF. Kromě věcné stránky se bude hodnotit i forma textu [**8 bodů**][1]. Buďte struční, avšak úplní. Váš postup popište tak, aby byl srozumitelný i čtenáři, který neabsolvoval předmět DVZ. Snažte se v textu odpovědět na všechny položené otázky, zdůvodňujte učiněná rozhodnutí. Text by měl obsahovat abstrakt, úvod, metodiku, výsledky a závěr. V **abstraktu** velmi stručně shrňte celý text. Jednou nebo pár větami vystihněte každou následující část textu. (Abstrakt se píše zpravidla až na závěr.) V **úvodu** stručně popište, čeho se váš projekt týká, co je vaším úkolem a co od své práce očekáváte. V **metodách** popište, jaká data máte k dispozici (včetně informace o tom kolik instancí a příznaků se v datech nachází, zda se v datech vyskytují chybějící a/nebo odlehlé hodnoty) a jakým způsobem budete postupovat (jaké metody použijete a proč). Diskutujte možné problémy, očekáváte-li nějaké, a navrhněte možná řešení. **Výsledky** by pak měly shrnout, k čemu jste dospěli aplikací zvolených metod na data. Vyberte relevantní výsledky, není třeba čtenáře zahlcovat přílišnými detaily či marginálními výsledky. Ve výsledcích by se měly objevit přesné odpovědi na všechny položené otázky, a to buď přímo v textu, nebo jako obrázek či tabulka. **Závěr** by měl diskutovat dosažené výsledky (co výsledky znamenají? jak je lze využít?) a stručně shrnout celý projekt.

Dbejte i na grafické zpracování. Z obrázků i tabulek by mělo být na první pohled jasné, co vyjadřují. V obrázcích nesmí chybět rozumný popis os (např. nikoli `attr1`, ale *hmotnost [kg]*). Jsou-li v obrázku textové popisy, měly by být jasné a čitelné. Označujete-li si pracovně příznaky např. jako `attr1`, `attr2`, nepatří tyto pracovní názvy rozhodně do publikačního výstupu. Na obrázky i

[1] Celkově tak můžete za vypracování semestrální práci získat 30 bodů (za její prezentaci pak dalších 10 bodů).

tabulky je dobré se odkázat z textu (pomocí čísel). Buďte konzistentní (zavedete-li nějaké termíny či zkratky, používejte je v celém textu; nemíchejte desetinné čárky a tečky).

V neposlední řadě se hodnotí i pravopisná správnost textu.