

# Zadání semestrální práce předmětu DVZ ZS 2012/13

## Téma: srdeční onemocnění

Koronární srdeční choroba (coronary heart disease (CHD)) je vážné onemocnění koronárních artérií, které se v důsledku různých vlivů zúží či zcela zablokují a nejsou tak schopny zásobovat srdce dostatečným množstvím kyslíku. Jedním z cílů tohoto projektu je proto identifikace faktorů ovlivňujících tuto nemoc. Na základě dat z retrospektivní studie všímající si příznaků naměřených na několika stovkách jihoafrických pacientů se budeme rovněž snažit chorobu diagnostikovat, tedy co nejpřesněji rozlišit pacienty od osob zdravých, a včasnou diagnostikou tak zvýšit šanci na úspěšnou léčbu.

Soubor `chd.csv` obsahuje data (matici pozorování), v souboru `chd.names` pak naleznete stručný popis jednotlivých příznaků.

Vášim úkolem je strojově klasifikovat pozorování (rozlišit pacienty od zdravých jedinců), modelovat míru obsahu tuku u sledovaných osob v závislosti na věku, a pokusit se identifikovat vnitřní strukturu dat.

Konkrétně:

### 1. Průzkumná analýza.

- Prohlédněte si data.
- 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 / vazby mezi příznaky / vazby mezi příznaky a typem nádoru.

### 2. Klasifikace osob na pacienty a nemocné CHD

- Vyberte alespoň dva klasifikátory vhodné pro tento typ dat. Podle čeho budete vybírat? Volbu zdůvodněte a fakticky podložte (obrázkem, číselně apod.).
- Na vhodné podmnožině příznaků vybrané klasifikátory natrénujte. Jaké úspěšnosti dosahují? Podle jakých příznaků se rozhodují?
- Jakou úspěšnost klasifikace očekáváte na nezávislých datech? Vyjádřete číselně. Mohli byste dát i nějakou záruku úspěšnosti takové budoucí klasifikace?
- Jaká je pravděpodobnost, že nově příchozí pacienta bude trpět CHD?
- Jaká je pravděpodobnost, že nově příchozí pacient bude klasifikován jako zdravý? Jaká bude naproti tomu pravděpodobnost, že nově příchozí zdravý jedinec bude klasifikován jako nemocný? 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 míry obsahu tuku (**adiposity**) na věku (**age**).

- Vhodnou metodou modelujte danou závislost a formálně ji zapište. Je závislost významná? Výsledky interpretujte.

### 4. Hledání podskupin v datech.

- Pokuste se v datech na základě hodnot příznaků (tj. bez zohlednění toho, zda se jedná o pacienta či zdravého jedince) nalézt vnitřní strukturu. Výsledky vizualizujte a interpretujte.
- Je možno dát nalezenou strukturu do souvislosti s rozlišením osob na nemocné a zdravé jedince? (Tj. nelze rozdělení na osoby nemocné a zdravé tušit již přímo v nalezené struktuře? Jinými slovy: neodpovídají nalezené podskupiny do značné míry pacientům/zdravým jedincům?) Vizualizujte a/nebo popište tabulkou.

Svojí práci shrňte ve formě vědeckého sdělení ve formátu PDF. Buďte struční, avšak úplní. Pište jen to, co sami považujete za důležité, avšak tak, aby byly vaše kroky jasné i čtenáři, který neabsolvoval předmět DVZ. Zdůvodňujte učiněná rozhodnutí. Buďte konzistentní (zavedete-li nějaké termíny, používejte je v celém textu; nemíchejte desetinné čárky a tečky).

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 na závěr.)

V **úvodu** popište, čeho se váš projekt týká, proč jej řešíte, co od něj 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 odlehle 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. 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 na grafické zpracování. Z obrázků i tabulek by mělo být na první pohled jasné, co vyjadřují. V obrázcích nesmí chybět popis os. Jsou-li v obrázku textové popisy, měly by být jasné a čitelné. Označujete-li si pracovní příznaky např. jako AT1, AT2 apod., nepatří tyto pracovní názvy rozhodně do publikačního výstupu. Na obrázky i tabulky je dobré se odkázat z textu (pomocí čísel).

V neposlední řadě se hodnotí i gramatická a stylistická správnost textu.