Úkol A – Explorace dat a lineární regrese – alcohol consumption

ZS 2016

Cíl

Cílem prvního úkolu je seznámit se s daty ze studie konzumace alkoholu studentů a pomocí lineární regrese zjistit, jaký je studijními výsledky a konzumací alkoholu o studentů.

Data

Data o konzumaci alkoholů studentů na nějaké portugalské střední škole, popisuje studenty z hlediska jejich úspěšnosti v oborech matematiky a portugalštiny, spolu parametry rodinného prostředí, času stráveného učením, cestováním do a ze školy, volným časem, konzumace alkoholu ve všední dny a o víkendech, apod. Konzumace alkoholu mezi studenty je velmi zajímavé téma a data studie mohou posloužit k objektivnímu posouzení jeho vlivu na studijní výsledky.

Explorace dat

Data ze studie o konzumaci alkoholu obsahují 33 příznaků, jejich stručný popis je uveden v tabulce 1 níže. V rámci explorace dat prozkoumejte příznaky věk, pohlaví, čas strávený cestováním, čas strávený učením, volný čas, spotřebu alkoholu ve všední dny a o víkendech. Seznamte s typem dat v jednotlivých příznacích, rozsahem jejich hodnot, jejich rozdělením, případně počtem chybějících a odlehlých hodnot. Zvažte, jestli některé z příznaků není potřeba transformovat pro potřeby dalších analýz. Prozkoumejte závislosti mezi jednotlivými příznaky a zajímavé závislosti vizualizujte.

Modelování pomocí lineární regrese

Pohled na konzumaci alkoholu u studentů je jednoznačně negativní. Pokuste na základě dat od úspěšných studentů (ti s nenulovým hodnocením na konci školního roku) zjistit, jaké proměnné mají skutečný vliv na úspěšnost studentů ve formě hodnocení na konci roku a jestli konzumace alkoholu skutečně negativně ovlivňuje studijní výsledky. Na základě exploračních vizualizací a modelů, rozhodněte které parametry jsou pro hodnocení na konci roku důležité a modelujte vztah mezi vybranými parametry a hodnocením na konci roku pomocí lineárního modelu. Výsledný model zhodnoťte z hlediska přesnosti modelu a jeho statistické významnosti.

Požadované kroky analýzy

Průzkumová analýza. [4 body]

- Dimenze dat (počet příznaků, počet instancí)
- Chybějící a odlehlé hodnoty, transformace dat
- Vizualizace

Modelování vztahu mezi konzumací alkoholu a studijními výsledky [6 bodů]

- Formální zápis modelu
- Interpretace koeficientů modelu
- Přesnost modelu
- Statistická významnost

Výsledky upravte do formy zprávy, která bude obsahovat stručný **úvod**, popis metod, které jste použili, v sekci **metody,** výsledky jejich aplikace na data v sekci **výsledky** a závěry, které jste zjistili interpretací výsledků v sekci **závěr**. Maximální délka zprávy je 3 stránky. Zprávu ve formátu pdf odevzdejte pomocí UploadSystemu.

school	škola, kterou student navštěvuje ('GP' - Gabriel Pereira nebo 'MS' - Mousinho da Silveira)
sex	studentovo pohlaví ('F' – female/žena nebo 'M' - malemuž)
age	studentův věk (od 15 do 22)
address	typ studentova bydliště ('U' – urban/městské nebo 'R' – rural/venkovské)
famsize	velikost rodiny ('LE3' - <=3 nebo 'GT3' - > 3)
Pstatus	soužití rodičů ('T' - living together/žijí spolu nebo 'A' – apart/rozvedeni)
Medu	vzdělání matky (0 - žádné, 1 – první stupeň ZŠ, 2 druuhý stupeň ZŠ, 3 SŠ or 4 VŠ)
Fedu	vzdělání otce
Mjob	zaměstnání matky
Fjob	zaměstnání otce
reason	důvod pro studium na dané škole
guardian	opatrovník
traveltime	doba cesty do školy (1 - <15 min., 2 - 15 až 30 min., 3 - 30 min. to 1 hod, nebo 4 - >1 hod)
studytime	týdenní doba studia(1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 až 10 hod, or 4 - >10 hod)
failures	počet předchozích nedokončení ročníku
schoolsup	podpora ze strany školy (yes nebo no)
famsup	vzdělávací podpora v rodině (yes nebo no)
paid	doučování (yes nebo no)
activities	mimoškolní aktivity (yes nebo no)
nursery	navštěvoval školku (yes nebo no)
higher	chce pokračovat na VŠ (yes nebo no)
internet	přístup k internetu doma (yes nebo no)
romantic	ve vztahu (yes neob no)
famrel	kvalita domácích vztahů (od 1 - very bad do 5 – excellent)
freetime	volný čas po škole (od 1 - very low do 5 - very high)
goout	chodí ven s kamarády (od 1 - very low do 5 - very high)
Dalc	konzumace alkoholu ve všední dny (od 1 - very low do 5 - very high)
Walc	konzumace alkoholu o víkendech (od 1 - very low do 5 - very high)
health	zdravotní stav (od 1 - very bad do 5 - very good)
absences	počet absencí (od 0 do 93)

Tabulka 1: Popis proměnných z studie konzumace alkoholu mezi studenty

Bonusová úloha – studijní stereotypy

Lineární regrese, nebo obecněji lineární model, je možné k testování řady poměrně složitých hypotéz. Lineární model může sloužit k porovnání závislostí na základě kategorických proměnných. Zajímavým příkladem v tomto může být vztah mezi pohlavím a hodnocením (nerovnost pohlaví, gender inequality). Lineární model lze použít k takovémuto porovnání. Lze na základě dat ze studie konzumace alkoholu rozhodnout, zda existují stereotypy jako že ‚muži jsou lepší v matematice', nebo ‚existuje nerovnost mezi pohlavími a projevuje se i v hodnocení studentů'?

Bonusová úloha umožňuje získání až 4 dalších bodů, ale vyžaduje samostudium teorie a praktického použití lineárních modelů nad rámec cvičení DVZ.