**VŠB – Technická univerzita Ostrava**

**Fakulta elektrotechniky a informatiky**

**Katedra informatiky**

METODY ANALÝZY DAT II

2017 Bc. Barbora Cigánková

Obsah

[1 Analýza dat 3](#_Toc481579474)

[1.1 Popis datasetu 3](#_Toc481579475)

[1.2 NA hodnoty 4](#_Toc481579476)

[1.3 Analýza 5](#_Toc481579477)

[1.3.1 Nahrazení NA nečastějšími hodnotami 6](#_Toc481579478)

[1.3.2 Predikce hodnot 7](#_Toc481579479)

[1.3.3 Zhodnocení 7](#_Toc481579480)

# Analýza dat

Pro analýzu dat byl zvolen dataset s názvem Wages and Education of Young Males dostupný ze ZIP souboru na webové stránce <http://vincentarelbundock.github.io/Rdatasets/>.

## Popis datasetu

Uvedený dataset obsahuje 4360 instancí s 12 vlastnostmi (+ ID instance). Jak již samotný název napovídá, jedná se o data týkající se mezd mladých mužů. Dataset obsahuje údaje z let 1980 – 1987 v rámci USA. Data byla získána v rámci národního longitudinálního výzkumu.

Dataset obsahuje 6 numerických atributů a 7 kategoriálních. Z kategoriálních atributů jsou 3 typu ano/ne, zbytek nabývá 3 – 12 hodnot. Dataset neobsahuje úplná data. Některé instance mají atribut, který nebyl vyplněn, resp. je označen jako NA. Instancí s neúplnými daty je v celém datasetu 1245. Všechny NA hodnoty jsou obsaženy v posledním atributu residence.

Názvy jednotlivých atributů jsou následující:

|  |  |
| --- | --- |
| * ID, | * married, |
| * nr, | * health, |
| * year, | * wage, |
| * school, | * industry, |
| * exper, | * occupation, |
| * union, | * residence. |
| * ethn, |  |

Prvním atributem je pořadové číslo instance. Druhým atributem je taktéž identifikační číslo, ale platné pro danou osobu v rámci výzkumu. I když je tedy počet instancí v celém datasetu 4360, osob zapojených ve výzkumu bylo 545. Pro každou osobu je tedy v datasetu 8 instancí (1 instance/ 1 rok).

Následujícím atributem je rok pořízení dat dané instance. Tento atribut, jak již bylo napsáno výše, nabývá hodnot 1980 – 1987. Atribut school obsahuje údaj o počtu let strávených ve škole danou osobou. Jelikož představuje v datasetu 8, po sobě jdoucích, instancí 1 osobu výzkumu, taktéž 8, po sobě jdoucích, instancí má stejný údaj atributu school. Tento fakt platí taktéž pro atribut ethn, který obsahuje údaj o etnické příslušnosti dané osoby a nabývá 3 hodnot – black, hisp a other.

Atribut exper představuje údaj o počtu let zkušeností. Údaj se tedy s rostoucím atributem year také zvyšuje. Atribut union je jedním z atributů, které nabývají hodnot typu ano/ne. Tento atribut zastupuje informaci o tom, zda měla daná osoba nastavený plat kolektivní smlouvou nebo nikoliv. Stejně tak atributy married a health, jsou stejného typu jako předchozí vlastnost. První jmenovaný, již podle názvu, zastupuje informaci o tom, zda je muž ženatý, druhý obsahuje informaci, zda má muž zdravotní problémy.

Desátým atributem (wage) je údaj o logaritmu hodinové mzdy dané instance. Atribut industry je kategoriální atribut nabývající největšího počtu hodnot, 12. Obsahuje informaci o odvětví průmyslu, ve kterém je instance zaměstnaná. Atribut occupation je taktéž kategoriálním atributem, který nabývá 9 hodnot a obsahuje informaci o konkrétním zaměstnání. Posledním atributem je atribut residence. Ten nabývá 4 hodnot – rural area, north east, nothern central, south. Je to tedy informace týkající se bydliště. Residence je jediným atributem, který „nabývá hodnot“ NA.

## NA hodnoty

Před jakoukoliv analýzou bylo nejprve nutné zamyslet se nad tím, jak naložit s NA hodnotami posledního atributu datasetu. Technik práce s chybějícími hodnotami v datech je několik. Výběr některé z nich závisí, mimo jiné, na povaze těchto dat, důvodu nevyplnění dat, ale také například na počtu chybějících hodnot.

Prvním typem metod je vymazání instancí s chybějícími hodnotami či jejich určité části. I když tato metoda vede ke zjednodušení dat, v případech datasetů s velkým množství takovýchto instancí by tato metoda mohla způsobit získání nepřesných statistických údajů. Možnou modifikací této metody je vymazání atributu, který chybějící data obsahuje. Tato modifikace je však možná pouze v případech, kdy se takto chybějící data objevují pouze v jednom atributů či ve velmi malém počtu atributů (vzhledem k jejich počtu). Další možným případem je situace, kdy daný atribut/y není potřeba do analýzy zahrnout.

Druhým typem je nahrazení chybějících hodnot. Poměrně častou technikou je doplnění chybějících dat průměrem či modem hodnot daného atributu. Nevýhodou této techniky je snížení variability dat. Druhým způsobem je nahrazení hodnot nejčastější hodnotou tohoto atributu (stejné nevýhody jako u průměru). Další možností je určitá predikce možné hodnoty těchto dat na základě ostatních atributů.

Pokud se chybějící data týkají kategoriálního atributu, je také možné považovat chybějící data za další hodnotu (skupinu), které atribut může nabývat.

Jelikož se chybějící data objevují bezmála ve třetině instancí analyzovaného datasetu, vymazáním všech těchto instancí bychom přišli o velké množství dat. Nicméně, chybějící hodnoty se objevují pouze v jednom atributu, a tedy možnost vymazání tohoto atributu by mohla přicházet v úvahu.

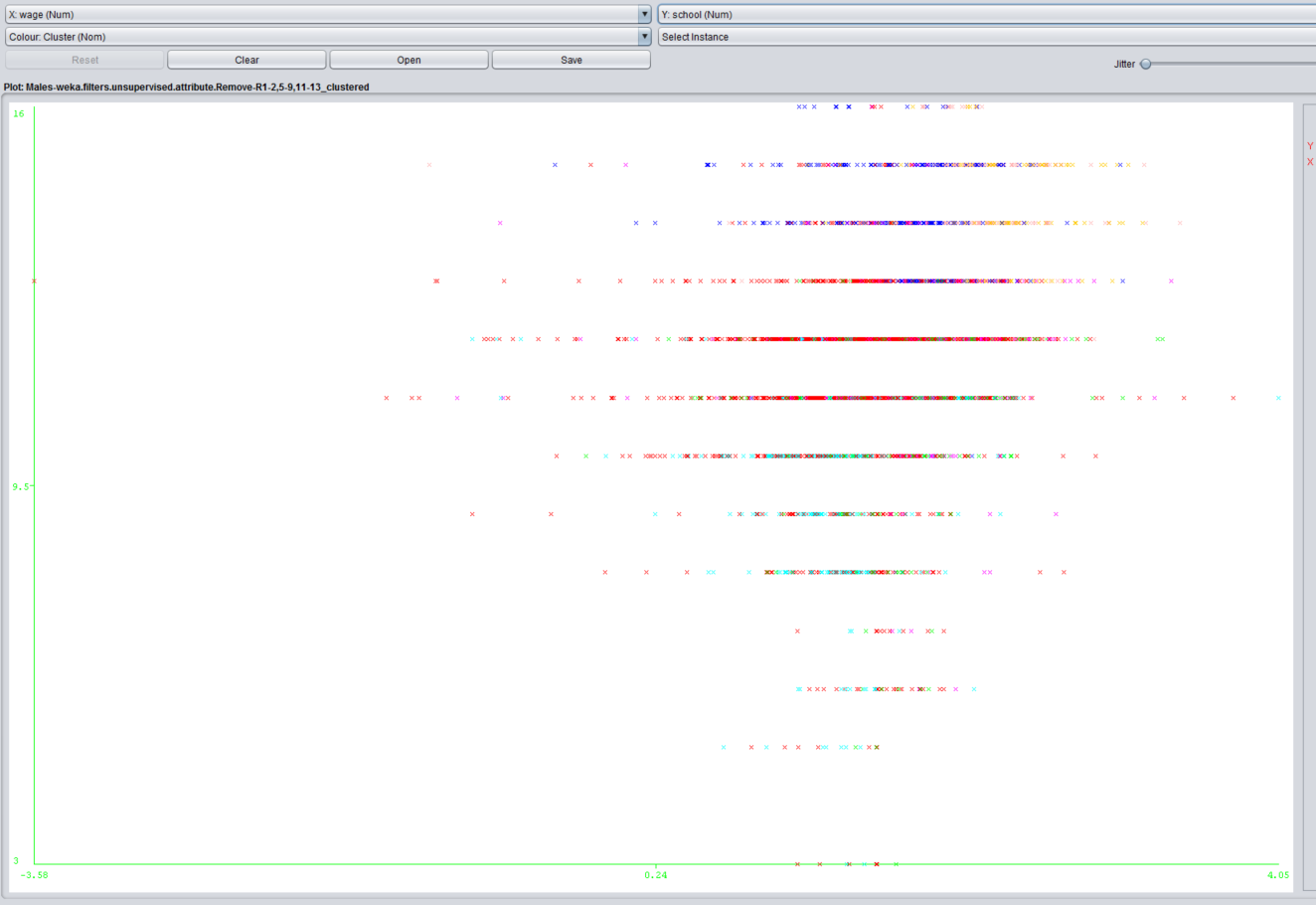
O něco vhodnější je druhý typ technik. Chybějící data mohou být nahrazena buď nejfrekventovanější hodnotou atributu nebo by mohla být odvozena z ostatních atributů.

Na data tedy aplikujeme tyto dvě metody – nahrazení chybějících dat nejčastější hodnotou, predikce hodnot podle zbylých atributů – provedeme analýzu takto předzpracovaných dat a vyhodnotíme výsledku jednotlivých způsobů.

## Analýza

Vybraný dataset bude analyzován postupně 2 metodami.

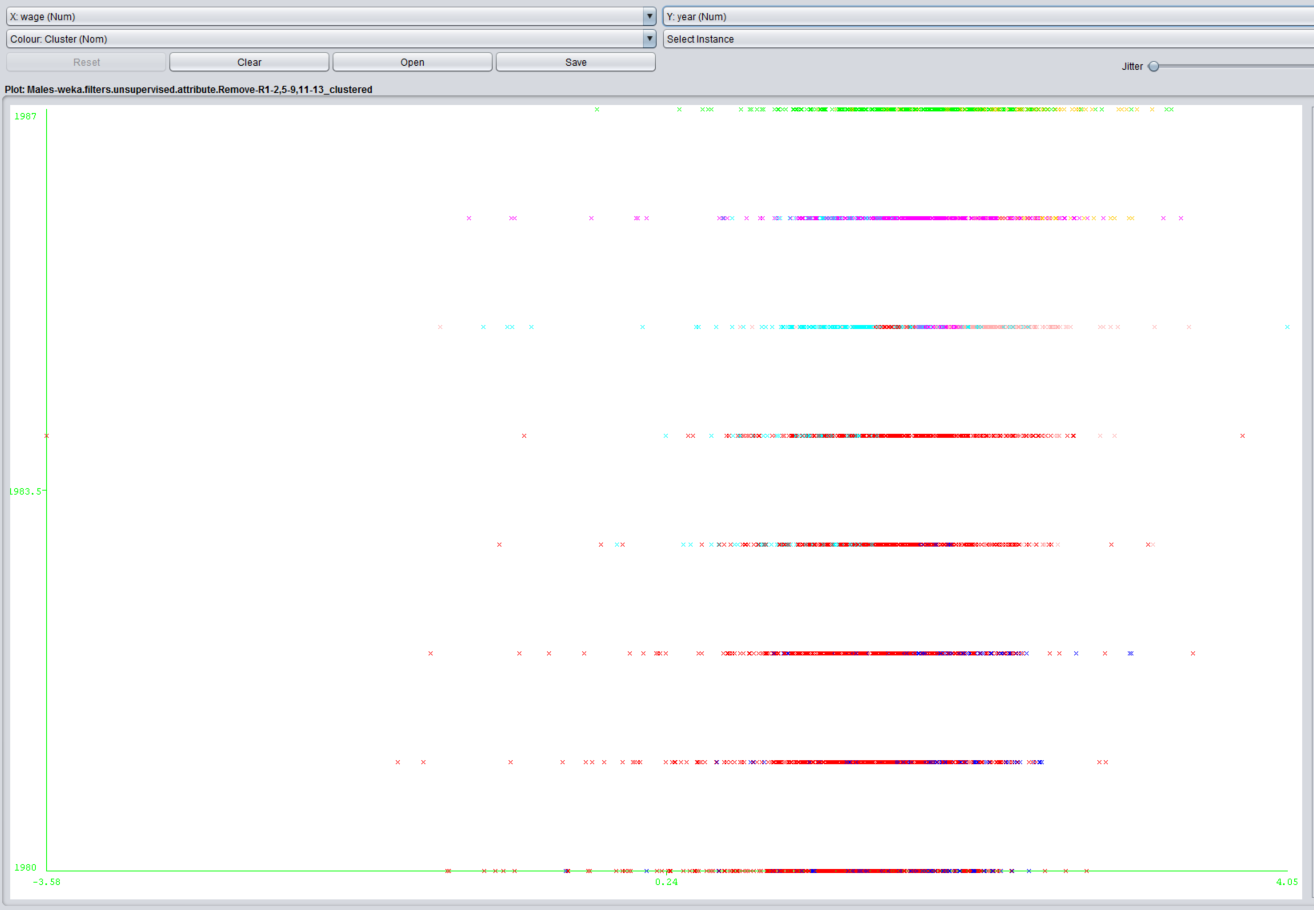
Metoda k-means (s Manhattan vzdáleností) bude použita pro shlukování dat s atributy year, school a wage. Podle utvořených shluků můžeme určit, zda je nějaká souvislost mezi těmito 3 atributy. K-means byl spuštěn v nástroji Weka.

Následující obrázek ukazuje výsledek k-means, kde na souřadnici x je zobrazen atribut wage a na souřadnici y atribut school.

Obrázek 1 K-means s Manhattan vzdáleností pro atributy year, school a wage (osa x - atribut wage, osa y - atribut school)

I když bychom předpokládali, že se v průběhu let bude klást větší důraz na vzdělání, a tedy delší vzdělání bude lépe ohodnoceno platem, Obrázek 1 ukazuje, že tomu tak není. Nicméně můžeme vidět, že v datasetu bylo více instancí s delším vzděláním než 10 let.

Druhým výsledkem tohoto k-means je porovnání atributu wage vzhledem k atributu year. Zde máme možnost pozorovat, zda se výplata zvyšovala v průběhu daného období, společně v závislosti na době vzdělání. Ani v tomto případě, však jasné výsledky nejsou, i když mírný posun, najčastěji se vyskytujících hodnot výplaty a nejvyšší výplaty, směrem doprava pozorovat můžeme. Na Obrázek 2 můžeme vidět, že v každém roce jsou v datasetu zastoupeny instance s prakticky s celou škálou atributu wage. Posun by byl pravděpodobně výraznější kdybychom měli k dispozici data za delší období.

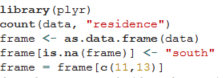


Obrázek 2 K-means s Manhattan vzdáleností pro atributy year, wage a school (osa x – atribut wage, osa y – atribut year)

Druhou použitou metodou bude Naivní Bayesův klasifikátor aplikovaný v jazyce R. Tato metoda bude aplikována na atributy residence a industry se snahou predikovat atribut ethn. Podle této klasifikace můžeme určit, zda na základě bydliště a odvětví průmyslu dokážeme podle dat určit etnickou příslušnost instance. Tato metoda bude provedena 2krát, s různě zpracovaným atributem residence (viz. NA hodnoty).

### Nahrazení NA nečastějšími hodnotami

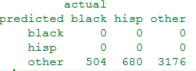
Nejprve zjistíme nejčastější hodnotu atributu residence pomocí package plyr a metody count. Následně nahradíme všechna NA atributu touto nejčastější hodnotou a vybereme ze všech dat pouze potřebné atributy, tedy industry (sloupec 11) a residence (sloupec 13).



Následně z balíčku e1071 použijeme metodu naiveBayes, kde parametry metody jsou data a predikovaný atrtibut. Po naučení klasifikátorů získáme výsledek predikce a vypíšeme jej do tabulky. V tomto případě byl výsledek predikce nic neříkající. Důvody mohou být 2. Buď jsou data natolik podobná, že klasifikátor doopravdy není schopen predikovat etnickou příslušnost nebo byl výsledek výrazně ovlivněn způsobem předzpracování NA hodnot. Ke kterému důvod se přiklonit uvidíme po aplikování druhého způsobu práce s NA hodnotami.

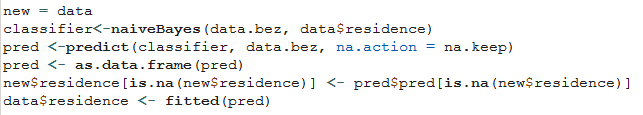


Obrázek 3 Naive Bayes v R

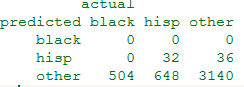


Obrázek 4 Naive Bayes při nahrazení NA hodnot nejčetnější hodnotou atributu residence

### Predikce hodnot

Predikce chybějících hodnot proběhla v jazyce R pomocí Bayese. Tyto predikované hodnoty byly následně doplněny do ostatních dat a byla provedena stejná predikce jako v předchozím případě. Stejně jako v předešlém případě je výsledek klasifikace sociálně uspokojivý, jelikož klasifikátor nedokázal klasifikovat instance do etnické příslušnosti.

Obrázek 5 Predikce NA hodnot atributu residence



Obrázek 6 Naive Bayes při nahrazení NA hodnot predikovanými hodnotami

### Zhodnocení

Ani v jednom případě předzpracování NA dat nedosáhl Naive Bayes takových výsledků, že bychom mohli s jistotou určit etickou příslušnost instance.