**Zadání**

**Tým:** Kateřina Fořtová (xforto00), Vojtěch Wawreczka (xwawre00)

**Vybraná datová sada:** Data o zaměstnanosti absolventů vysokých škol

**Popis dat:** Daná datová sada se zabývá analýzou vybraných uchazečů o práci na jejich akademické půdě. Jedná se o způsob, jak získat nabídku na stáže a celkově první pracovní zkušenosti. Data berou v potaz výsledky v průběhu studia, pracovní zkušenosti, zvolenou specializaci a také obsahují údaje o rozhodnutí, zda byl vysokoškolský uchazeč na danou např. stáž přijat a jeho finanční ohodnocení.

Daná datová sada obsahuje jediný .csv soubor s následujícími atributy:

**sl\_no** - identifikační číslo

**gender** - pohlaví

**ssc\_p** - percentil při 10. ročníku školy (středoškolské vzdělání)

**ssc\_b** - místo vzdělání při 10. ročníku školy (středoškolské vzdělání)

**hsc\_p** - percentil při 12. ročníku školy (vyšší středoškolské vzdělání)

**hsc\_b** - místo vzdělání při 12. ročníku školy (vyšší středoškolské vzdělání)

**hsc\_s** - zvolená specializace při 12. ročníku školy (vyšší středoškolské vzdělání)

**degree\_p** - percentil při vysoké škole (bakalářské vzdělání)

**degree\_t** - zvolený obor vysokoškolského vzdělávání (bakalářského)

**workex** - pracovní zkušenost (ano / ne)

**etest\_p** - výsledek testu zaměstnatelnosti (v procentech)

**specialisation -** zvolený obor vysokoškolského vzdělávání (magisterského)

**mba\_p -** percentil při vysoké škole (magisterské vzdělání)

**status** - přijetí uchazeče (ano / ne)

**salary** - příjem

Data bude nutné předzpracovat před samotnými dolovacími úlohami. V první řadě se bude jednat o odstranění sloupce **sl\_no**, který obsahuje pouhý číselný identifikátor uchazeče a nehraje roli při dolovací úloze. Pole **salary** obsahuje několik null atributů, které udávají nepřijaté uchazeče, těmto polím nemůžeme přiřadit např. průměrný plat na základě ostatních uchazečů, protože by to bylo konfliktní s kontextem dat, protože nepřijatí uchazeči plat nedostanou. Tyto null atributy tedy budou nahrazeny číslem 0, kdy se jedná o nepřijaté uchazeče.

**Zvolené úlohy**

**Klasifikace pohlaví uchazeče na základě ostatních atributů**

V této úloze se chceme zaměřit na klasifikaci pohlaví uchazeče na základě různých modelů, které jsou dostupné v prostředí Rapid Miner - např. Naive Bayes, Deep Learning nebo i dalších dostupných. Chceme tedy získat znalosti o tom, zda je možné právě z údajů od studijních výsledků, specializace a dalších, až po konečné přijetí či nepřijetí uchazeče odhadnout právě jeho pohlaví.

**Predikce přijetí uchazeče na základě pouze některých zvolených atributů**

Soubor dat obsahuje atribut **status**, který určuje, zda byli daní uchazeči přijati na pracovní pozici. Cílem tohoto úkolu je předpovědět pouze na základě pohlaví uchazeče a získaných percentilů napříč vzděláváním (**ssc\_p**, **hsc\_p**, **degree\_p**, **etest\_p**, **mba\_p**), zda byl vybrán. Opět chceme využít více modelů dostupných v nástroji Rapid Miner. Získané znalosti by byly nápomocné k rozhodnutí, zda lze pouze z pohlaví a studijních výsledků opravdu předpokládat výsledek studenta při pohovoru a nebo nikoliv.

**Shlukování na základě dosažených percentilů napříč studiem**

Pro shlukování plánujeme využít metodu K means. Studenti budou rozřazeni do např. tří shluků na základě všech dosažených percentilů (**ssc\_p**, **hsc\_p**, **degree\_p**, **etest\_p**, **mba\_p)**. Normalizace dat by neměla být nutná, protože u všech údajů se jedná o jednotné percentilové vyjádření. U každého shluku budeme moct vyčíst průměrné údaje pro každý atribut studentů v daném shluku zastoupených. Můžeme tedy zjistit, zda budou mít studenti v každém se shluků vždy všechny percentily lepší ve srovnání s jiným shlukem, nebo nikoliv. Také můžeme následně vypočítat průměrný plat každého shluku a zjistit tak, zda studenti s vyššími percentily dostávají i vyšší plat.