Projekt do předmětu ZZN

Řešení úlohy

| **Vybraná datová sada:** | Data o zaměstnanosti absolventů vysokých škol |
| --- | --- |
| **Řešitelé:** | Kateřina Fořtová (xforto00)  Vojtěch Wawreczka (xwawre00) |
| **Datum:** | 5.12.2021 |

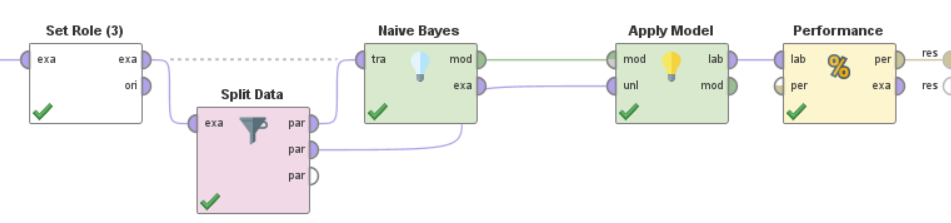
## Předzpracování dat

Data byla nutna před samotnými dotazy předzpracovat. Na načtená data byl aplikován blok Select Attributes, který vybral atributy, se kterými se následně bude pracovat - atribut sl\_no obsahuje pouze číselný identifikátor uchazeče, nehraje roli při dolovací úloze a proto byl nevybrán. Blok Replace Missing Values následně nahrazuje všechny NULL hodnoty atributu salary číselnou hodnotou 0.



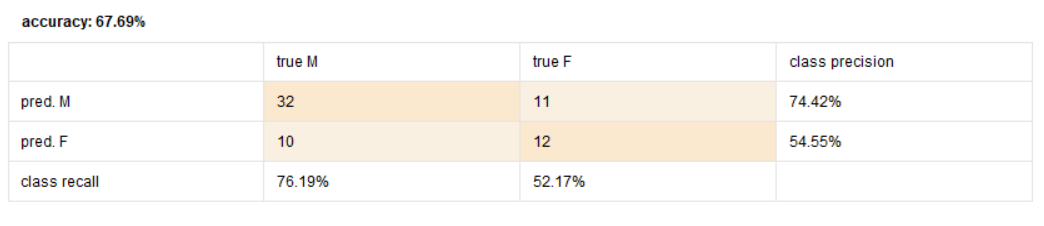
## Klasifikace pohlaví uchazeče

Tato úloha je zaměřena na klasifikaci pohlaví uchazeče na základě různých modelů, které jsou v prostředí Rapid Miner dostupné (Naive Bayes, Deep Learning apod.). Získané znalosti nám pomohou díky informacím o studiu a úspěšnosti při hledání zaměstnání s určitou pravděpodobností určit pohlaví daného studenta.

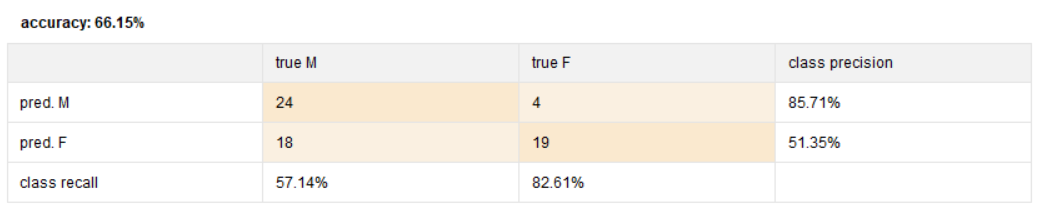


Při klasifikaci bylo využito bloku Set Role, který nám atribut gender nastavil jako label - třídu, kterou se bude model snažit predikovat. Data jsou následně rozdělena na trénovací a testovací sadu. Na blokovém schématu výše je uveden model Naive Bayes, avšak bylo experimentováno i s dalšími modely, které prostředí Rapid Miner nabízí. Natrénovaný model byl aplikován na testovací data a byla změřena přesnost klasifikace dle pohlaví pomocí bloku Performance.

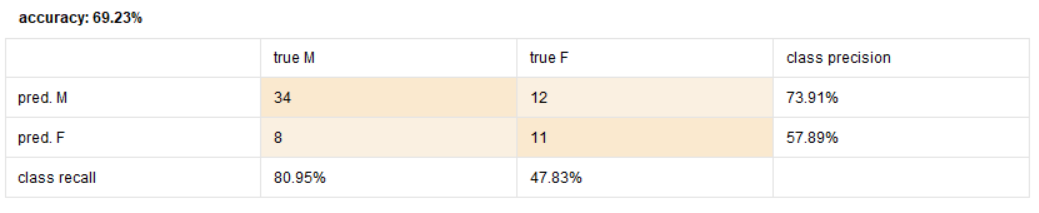
Klasifikace pomocí Naive Bayes metody:



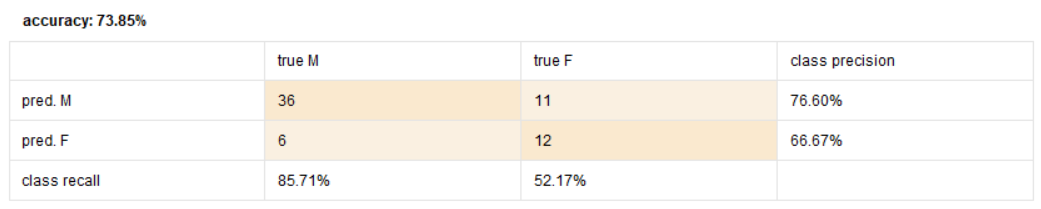
Klasifikace pomocí Deep Learning při předdefinovaném nastavení modelu (activation - Rectifier, epochs - 10):



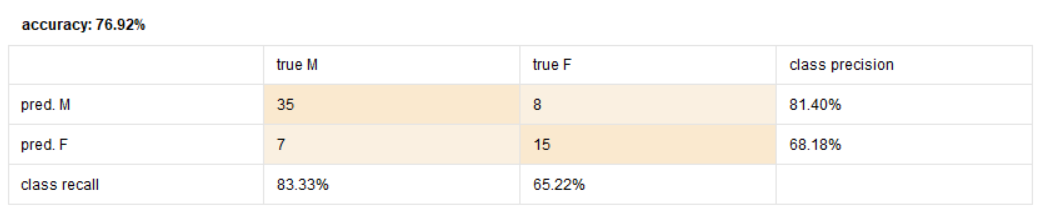
Klasifikace pomocí Deep Learning při nastavení modelu (activation - Tanh, epochs - 1000):



Klasifikace pomocí Random Forest při předdefinovaném nastavení modelu (number of trees - 100, maximal depth - 10):



Klasifikace pomocí Random Forest při nastavení modelu (number of trees - 150, maximal depth - 30):

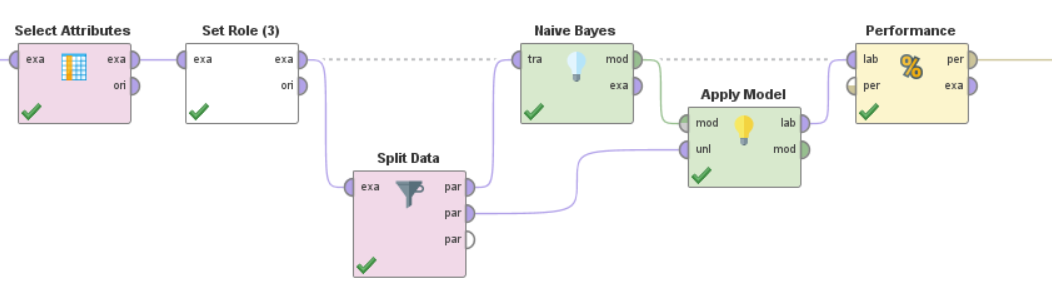


Zhodnocení:

Při klasifikaci pohlaví na základě všech atributů bylo dosaženo nejvyšší přesnosti 76.92 % při klasifikaci pomocí modelu Random Forest. Narozdíl od modelů Deep Learning a Naive Bayes se nám u tohoto modelu pro predikci podařilo překročit přesnost 70 %.

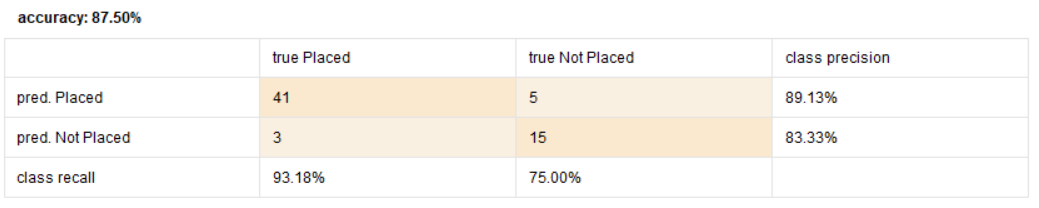
## Predikce přijetí uchazeče na základě zvolených atributů

Cílem druhé úlohy je predikce přijetí uchazeče o zaměstnání na základě jeho pohlaví a úspěšnosti studia (procenta úspěšnosti napříč vzděláváním a výsledku testu zaměstnatelnosti). Pomocí různých modelů z prostředí Rapid Miner získáme znalosti potřebné k tomu, abychom byli schopni predikovat výsledek absolventa školy při pracovním pohovoru.

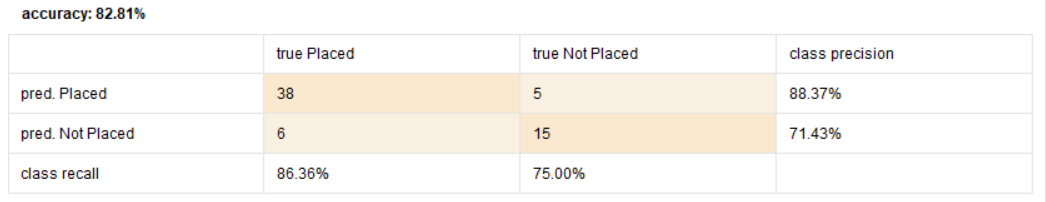


Blokové schéma predikce přijetí uchazeče je podobné schématu předchozí úlohy, avšak při Select Atrributes jsou vybrána pouze pole s atributy status, *gender*, *ssc\_p*, *hsc\_p*, *degree\_p*, *etest\_p* a *mba\_p* a jako label je nastaven atribut status.

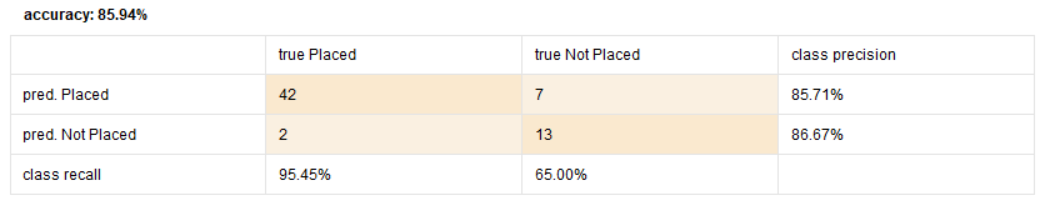
Klasifikace pomocí Naive Bayes metody:



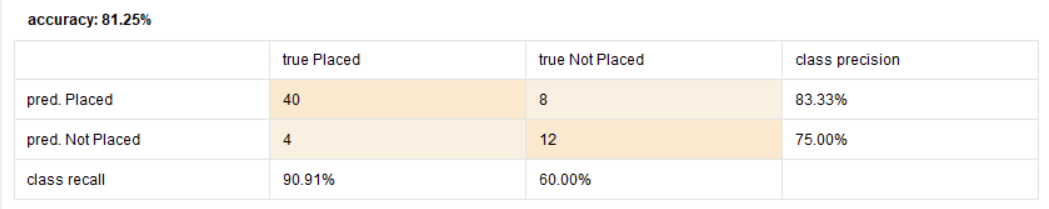
Klasifikace pomocí Deep Learning při předdefinovaném nastavení modelu (activation - Rectifier, epochs - 10):



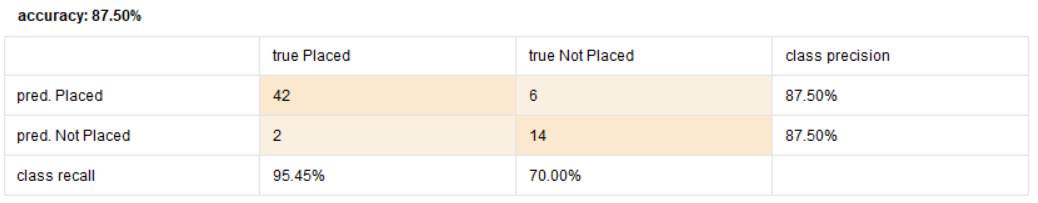
Klasifikace pomocí Deep Learning při nastavení modelu (activation -Tanh, epochs - 1000):



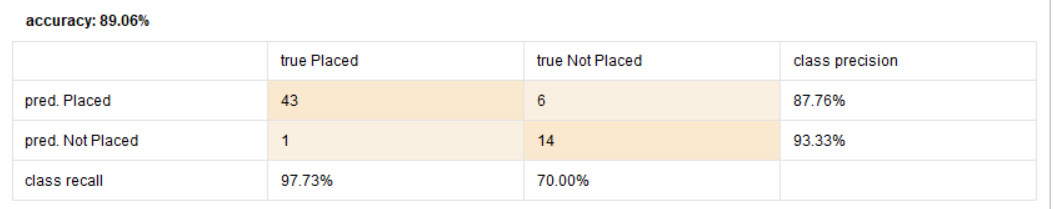
Klasifikace pomocí Deep Learning při nastavení modelu (activation - Rectifier, epochs - 1000):



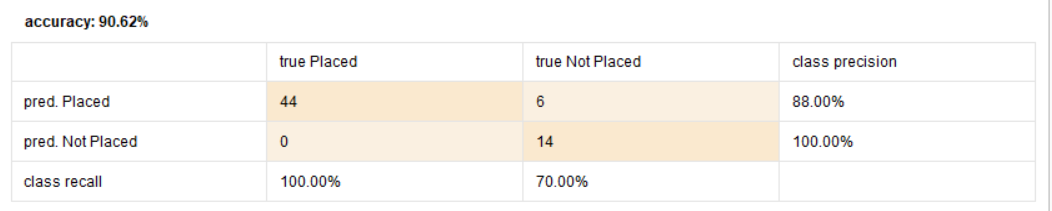
Klasifikace pomocí Random Forest při předdefinovaném nastavení modelu (number of trees - 100, maximal depth - 10):



Klasifikace pomocí Random Forest při nastavení modelu (number of trees - 150, maximal depth - 30):



Klasifikace pomocí Random Forest při nastavení modelu (number of trees - 200, maximal depth - 50):



Zhodnocení:

Při klasifikaci pohlaví na základě dosažených úspěšností v průběžných studijních výsledcích se nám podařilo dosáhnout vyšší úspěšnosti jak v předchozí úloze. Nejúspěšnějším modelem byl opět Random Forest, u kterého se nám úpravami parametrů podařilo dosáhnout úspěšnosti 90,62 %. I ostatní metody však dokázaly dosáhnout přesnosti více jak 80 %.

## Shlukování na základě dosažených procent úspěšnosti napříč studiem

Poslední úloha se bude zabývat odpovědí na otázku, zda lze studenty objektivně dělit do skupin podle úspěšnosti a zda úspěšnosti studenta při studiu odpovídá i jeho následný plat v zaměstnání. Pro tuto analýzu využijeme princip shlukování, konkrétně metodu K-means. Studenti tak budou rozděleni do několika shluků. Mezi jednotlivými shluky budeme následně porovnávat výše procent úspěšnosti i výše průměrných platů v rámci shluku.

Z datové sady využijeme pro rozdělení do shluků atributy *ssc\_p*, *hsc\_p*, *degree\_p*, *etest\_p* a *mba\_p*. Normalizace těchto atributů není nutná, protože všechny jsou vyjádřeny pomocí procent. Pro následné porovnání shluků budeme potřebovat také atribut *salary*.