## Projekt ADD - MovieLens

Mariusz Hausenplas s10638

Zbigniew Frysztak s9483

Piotr Bryński s10463

Do wykonania projektu użyliśmy danych pochodzących z serwisu **MovieLens**. W projekcie zrealizowanym w programie **RapidMiner** zaimportowane zostały następujące **zbiory danych**:

1. **Movies** - zbiór 9124 rekordów z opisywanymi filmami
2. **Ratings** - zbiór 100004 rekordów przechowujących oceny powyższych filmów
3. **Tags** - zbiór 1296 rekordów z “tagami”

Po zaimportowaniu danych do skryptu na zbiorze **Movies** zastosowany został operator “**Generate Attributes**”, w którym wygenerowane zostały atrybuty na podstawie już istniejących. Operacja miała na celu ‘przycięcie’ wartości i dostosowanie ich do efektywniejszej obróbki.

Należy nadmienić że w zbiorach **Ratings** oraz **Tags**, dla niektórych rekordów powtarzały się wartości. W celu wyeliminowania ich, dla poprawy działania skryptu, oba zbiory danych zostały przetworzone za pomocą skryptu w Javie.

Następnie wszystkie zbiory zostały ze sobą złączone za pomocą operatora “**Join**”

Po złączeniu, za pomocą operatora “**Filter Examples**”, ze zbiorów odflitrowane zostały wszystkie rekordy, których parametr ‘rating’ był brakujący. W tak zmodyfikowanych danych wartości numeryczne zostały następnie zastąpione ich odpowiednikami typu “**Polynominal**”

Do predykcji wartości “**ratings**” zostały użyte dwie metody, różnią się one zastosowanymi klasyfikatorami i w celu określenia najwydajniejszego sposobu predykcji ich wyniki będą porównywane na podstawie danych zebranych z wykonania procesu z użyciem parametrów o różnych wartościach.

1. Klasyfikator “**Naive Bayes**” z zaokrągleniem

W tym przypadku predykcja odbywa się z pomocą klasyfikatora dającego wyniki w zakresie od 0 do 5, czyli zgodnie z zakresem faktycznych ocen. Niestety na chwilę obecną proces musi zostać odpowiednio “dostrojony” by uzyskać większy stopień zgodności predykcji z faktycznymi ocenami filmu. Współczynnik “**accuracy**” operatora “**Classification Performance**” wyniósł 22.62%

b) Klasyfikator “**Naive Bayes**” z operatorem “**Numerical to Binominal**”

W tym przypadku predykcja jest “mocno zaokrąglona” ponieważ takie ustawienie procesu skutkuje wynikami z zakresu 0 i 1. W momencie gdy przewidywana wartość zawiera się w zbiorze od 0 do 2,5 zwracana jest wartość false oraz od 2,5 do 5 zwracana jest wartość true. W tym przypadku współczynnik “**accuracy**” operatora “**Classification Performance**” wyniósł 64.14%. Jest to wynik większy od poprzedniego jednakże w kontekście zmniejszonej dokładności nie daje on w pełni satysfakcjonujących wyników.

Na chwilę obecną prototyp rozwiązania pozwala uzyskać predykcję, jednakże na niekoniecznie precyzyjnym poziomie. Kolejnym krokiem będzie dopracowanie procesu i dostosowanie parametrów tak by dawał on wynik na poziomie co najmniej od 50% wzwyż.