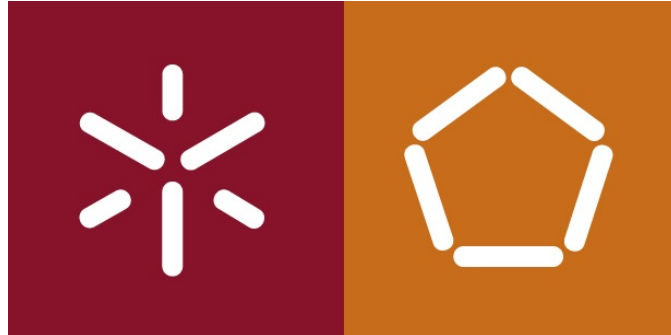


UNIVERSIDADE DO MINHO

MESTRADO INTEGRADO EM ENGENHARIA INFORMÁTICA



Sistemas Baseados em Similaridade

RELATÓRIO DO TRABALHO PRÁTICO

CONCEÇÃO E IMPLEMENTAÇÃO DE UM SISTEMA DE RECOMENDAÇÃO

GRUPO 12



Nuno Silva
A78156



Rui Vieira
A74658

5 de Janeiro de 2020

Conteúdo

1	Introdução e Objectivos	1
2	Dataset	1
3	Tratamento Dados	1
4	Workflows	2
5	Sistemas Recomendação	6
5.1	Baseado em Clusters	7
5.2	Baseado em Regras Associativas	7
6	Resultados Obtidos	9
6.1	Baseado me Clusters	9
6.2	Baseado em Regras Associativas	11
7	Sugestões e Recomendações	13
8	Conclusão	14

1 Introdução e Objectivos

O desenvolvimento de um sistema de recomendação, tem como principal objetivo indicar de maneira mais precisa e robusta, produtos que vão ao encontro dos gostos e preferências do consumidor em questão.

O sistema criado, para este trabalho prático, é um que visa a recomendação de filmes, para tal foi implementada uma tipologia híbrida, formada por sistemas baseados em conteúdo e de regras associativas. Permitindo a visualização de recomendações que foram baseadas em características do próprio filme, ou em tendências registadas.

De forma a apresentar a melhor recomendação possível, é de extrema importância, um bom tratamento sobre o *dataset*, com o objetivo de tirar o melhor proveito da informação.

2 Dataset

Os sistemas recomendação criados, utilizam um dataset nomeado "*movie_metadata*", este dataset foi recolhido da plataforma *Kaggle*, ao qual era referente a uma competição. Possuindo dados, claro está, que caracterizam um filme tal como "Color", "Gross", "Year", "Aspect Ratio", "Imdb Score", "Language", etc.

Este dataset possui cerca de 5043 linhas e 28 colunas, sendo que não se trata de um dataset completo pois falta-lhe informação, perto de 32512 células vazias. Sendo que estas mesmas células e as restantes serão tratadas. Onde de seguida será explicado o tratamento aplicado.

3 Tratamento Dados

O tratamento dos dados foi realizado, após uma análise gráfica e à leitura do *CSV* que possui a informação utilizada. Este mesmo tratamento, é dividido em 2 processos, um principal (o de baixo) e um secundário (o de cima).

Para o processo secundário, primeiro de tudo a coluna "*Genres*" foi dividida, pois seguia o formato "género1|...", assim diversas tabelas, cada contendo um único registo sobre o género de cada filme. Estas mesmas colunas e células que sofrerão diversas junções de forma a obter-mos uma única coluna contendo todos os géneros cinematográficos registados. De seguida, no processos primário, é removida a linha 4447, uma linha completamente desformatada, provocando a criação de colunas sem qualquer finalidade, onde estas também serão removidas. De seguida, é criada um coluna por cada género onde é

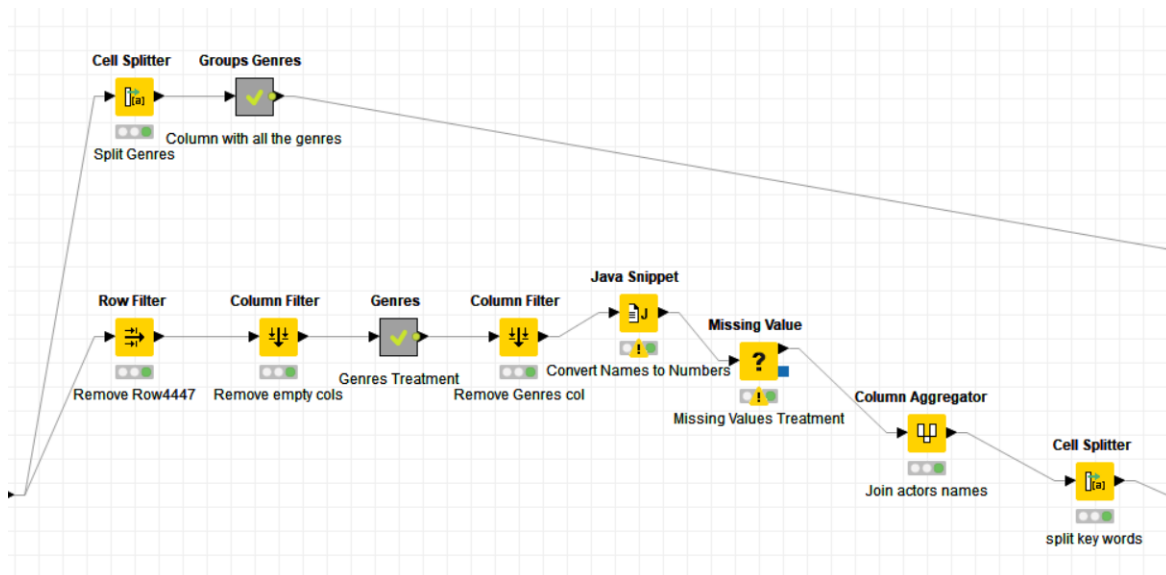


Figura 1: Tratamento utilizado sobre os dados

notado em cada célula o valor 1 caso o género esteja presente, caso contrário 0. A seguir são removidas todas as linhas que tenham em falta strings, por sua vez os valores inteiros e com decimais serão tratados por via de interpolação linear.

Por fim as tabelas contendo o nome dos atores foram agregadas, e a coluna contendo as palavras chave que caracterizam um filme, foi decomposta.

4 Workflows

O projeto criado é composto por 6 *workflows*, "Import Data", "Data Treatment", "Cluster Creation", "Data Representation", "Association Rules", e "Interface".

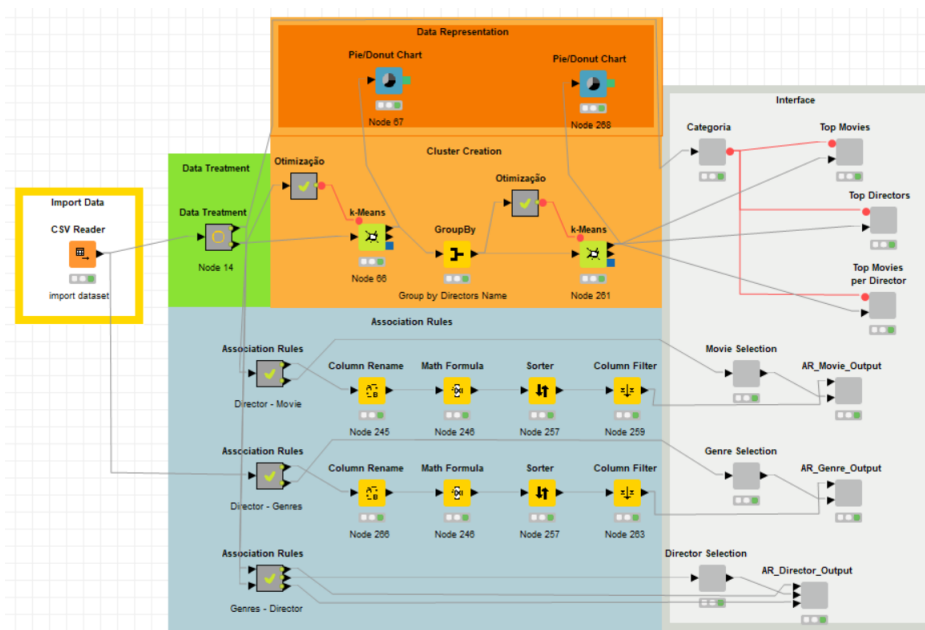


Figura 2: Visão geral sobre o projecto criado

Estes *workflows* criados, permitem uma melhor compreensão sobre a evolução do sistema, e em simultâneo facilitar a sua escalabilidade e atualização. O sistema, tal como referido na secção anterior, começa na leitura e tratamento do dataset utilizado. Onde o workflow "Data Treatment" fica encarregue da preparação dos dados, aos diferentes sistemas de recomendação implementados.

De seguida, os dados passam por um sistema de recomendação baseado em *Clusters*, fazendo uso do algoritmo *K-Means* de forma a agregar-mos os dados tratados previamente. O princípio base, é a segmentação dos dados com base no tipo de filme, caracterizado pelo seu género e o rating atribuído.

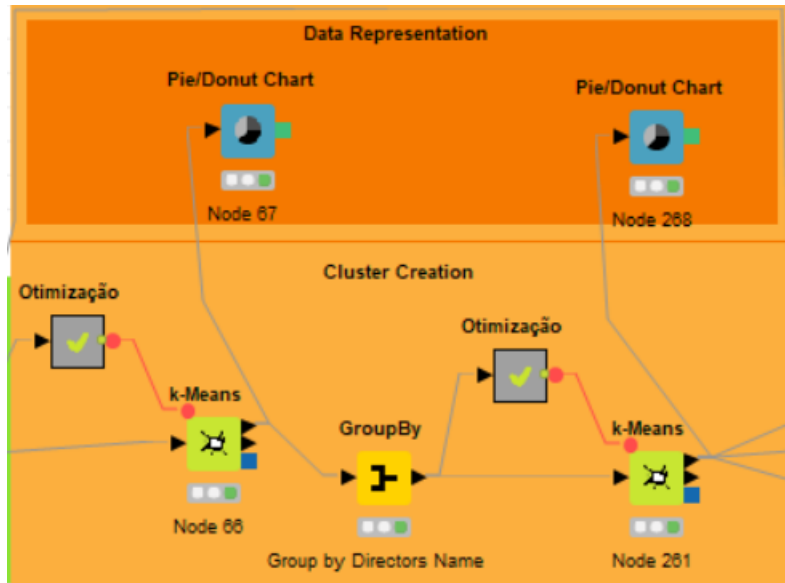


Figura 3: Workflows "Cluster Creation" e "Data Representation"

A recomendação é feita tendo em conta a preferência de um utilizador para com um género, tendo por base os dados recolhidos do cluster que mais se assemelha. Durante a utilização do K-Means, é necessário indicar previamente um número de clusters ideais, sobre o qual os dados serão segmentados. Este valor será previamente calculado e otimizado, utilizando o "método cotovelo".

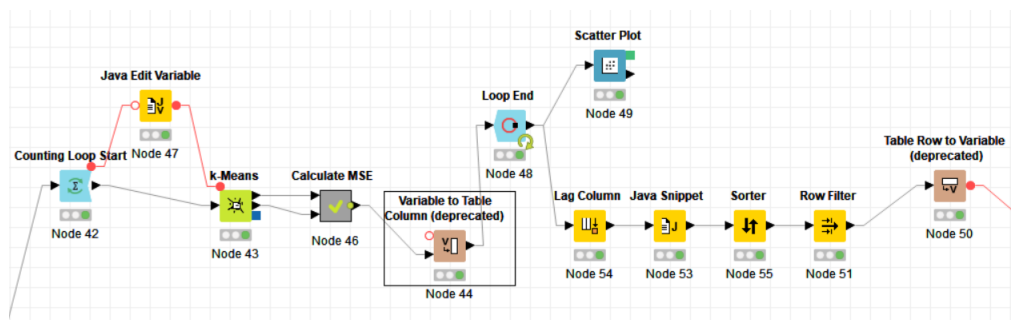


Figura 4: Método cotovelo

Em que constatamos que o número ideal de clusters será determinado, tendo em conta o MSE (Mean Squared Error).

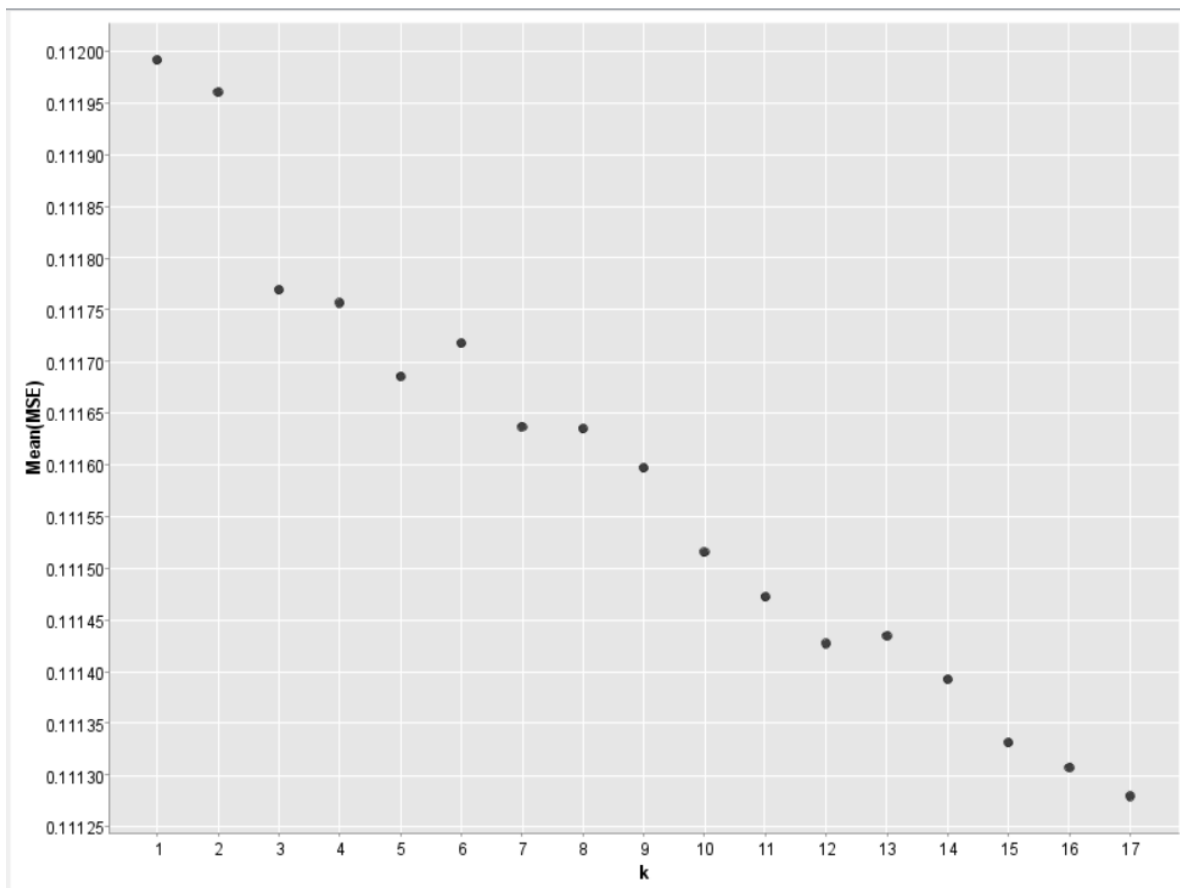


Figura 5: Visualização gráfica método cotovelo

Pelo método referido, constatamos que os dados serão segmentados utilizando 2 clusters, pois é entre 2 e 3 que ocorre uma descida mais acentuada no MSE.

A segunda componente do nosso projecto, corresponde ao sistema de recomendação baseado em regras associativas.

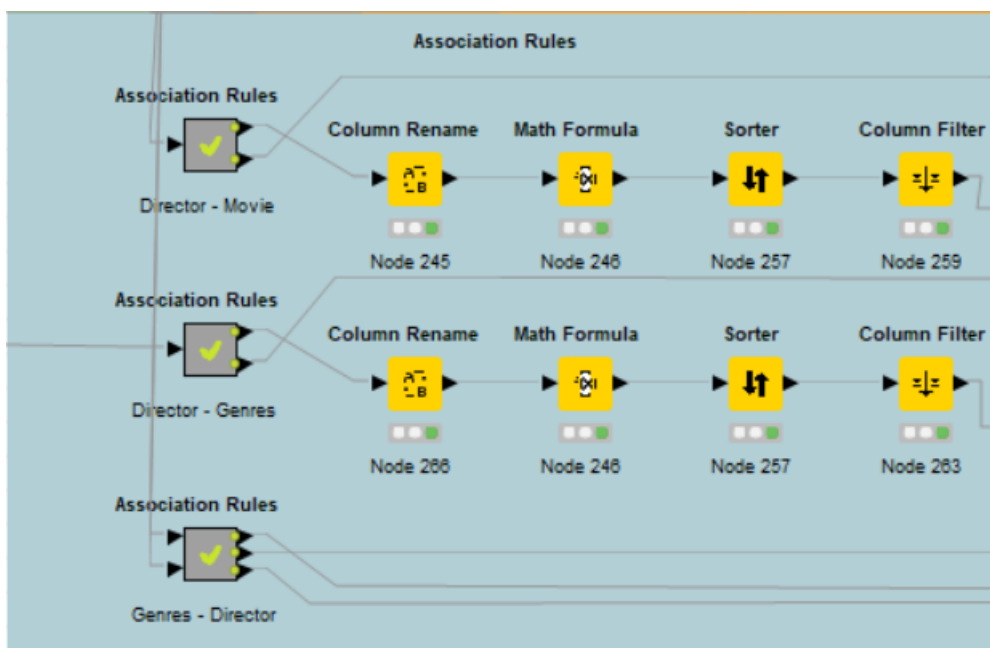


Figura 6: Workflow "Association Rules"

A construção das regras é feita em torno dos filmes e géneros que os realizadores

fizeram e uma outra com base nos géneros para obter os melhores directores. Para isso são utilizados os nodos "Association Rule Learner" e "Association Rule Learner(Borgelt)". A configuração para "Association Rule Learner(Borgelt)", foi escolhida de forma a obter o máximo número de regras com suporte mínimo de 1% e confiança de 45%. Por outro lado para "Association Rule Learner" foi selecionado um suporte de mínimo de 30% e uma confiança de 80%, de forma a obter regras com maior qualidade possível.

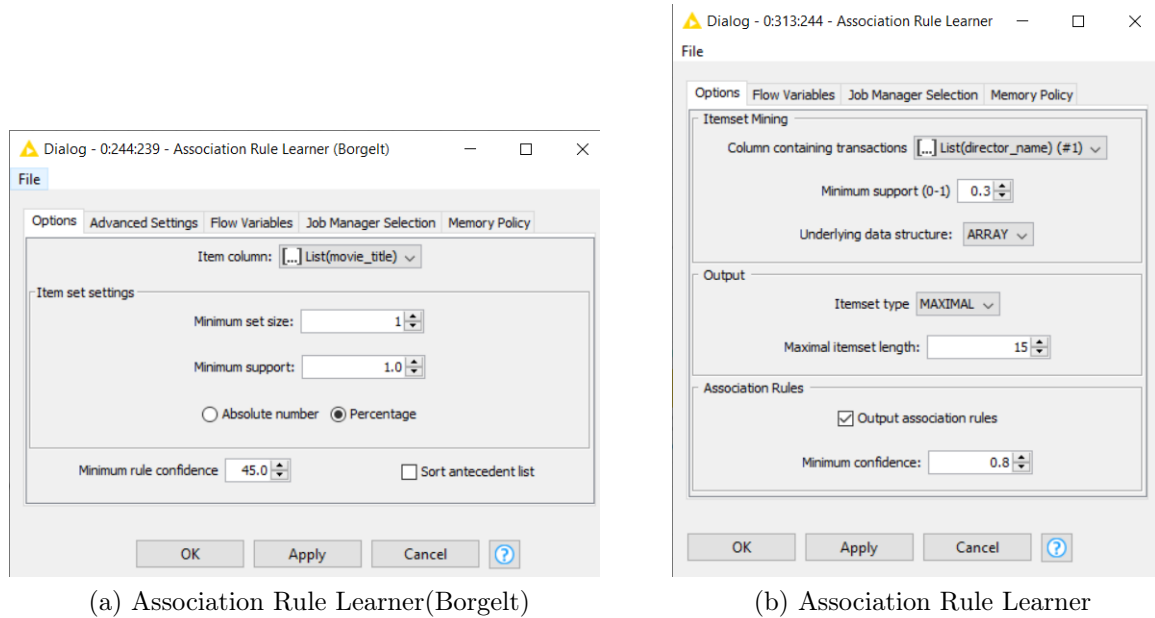


Figura 7: Configurações utilizadas

A ordenação das regras para Association Rule Learner(Borgelt), é feita através da multiplicação entre *ItemSetSupport* e *RuleConfidence*, sendo que para Association Rule Learner é feita pelo maior valor de *lift* encontrado e como critério de desempate a confiança.

D Support	D Confide...	D Lift	S Consequent	S implies	[...] Items
0.3	1	3.333	Rob Cohen	<---	[Richard Donner, Joel Schumacher...
0.3	1	3.333	Tarsem Singh	<---	[Richard Donner, Steven Spielber...
0.3	1	3.333	Ridley Scott	<---	[Martin Campbell, ?, James Mangol...
0.3	0.857	2.857	Jon Turteltaub	<---	[?, Chris Columbus]
0.3	0.857	2.857	Bryan Singer	<---	[Richard Donner, ?]
0.3	0.857	2.857	Luc Besson	<---	[Steven Spielberg, ?]
0.3	0.857	2.857	Stephen Sommers	<---	[Tim Burton, ?]

Figura 8: Excerto do output Association Rule Learner

Por fim, o workflow "Interface", é constituído por *wrapped metanodes*, em que alguns deles, "Categoria", "Movie Selection", "Genre Selection" e "Director Selection", possibilitam uma escolha por parte do utilizador, que posteriormente essas mesmas escolhas influenciam o output referente aquilo que é pretendido. Por exemplo, "Movie Selection" recebe uma lista de filmes, permitindo ao utilizador seleccionar alguns filmes que já

viu, para posteriormente, e fazendo uso das regras de associação, indicadas anteriormente, fazer a recomendação de novos filmes, pelo wrapped metanode *AR_Movie_Output*.

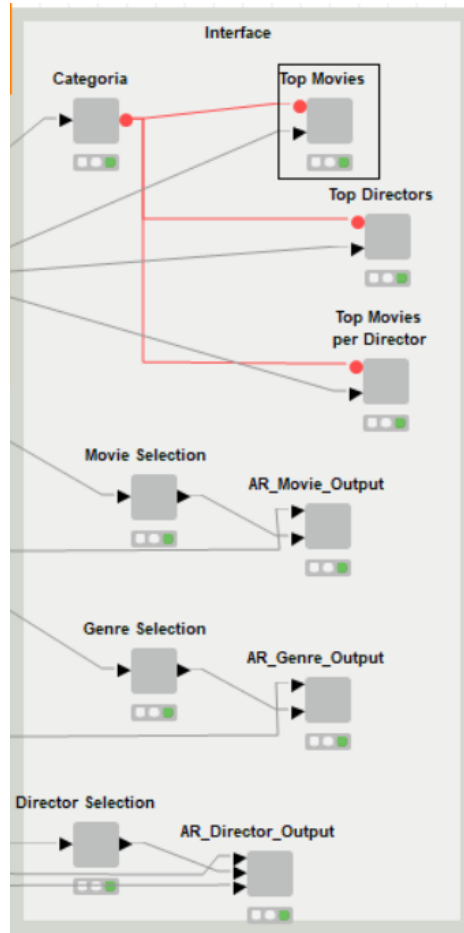


Figura 9: Workflow "Interface"

Após ser feita escolhas por parte do utilizador, temos os wrapped nodes com a função de transmitir o output contendo as respetivas recomendações, baseadas nas suas escolhas. O metanode *Top Movies*, apresenta ao utilizador o top 5 filmes com base no género, escolhido anteriormente no metanode *Categoria*, e na sua classificação no Imdb.

Já os wrapped metanodes *Top Directors* e *Top Movies per Director*, apresentam consoante a categoria selecionada, os melhores directores e as suas classificações médias, e uma lista dos melhores directores e os filmes que realizaram, respetivamente.

Os wrapped metanodes que começam por "*AR_*", são apresentados resultados com base nas regras de associação, referidas anteriormente.

5 Sistemas Recomendação

Para este trabalho prático forma criados 2 sistemas de recomendação individuais, estando integrados num sistema recomendação híbrido.

5.1 Baseado em Clusters

Este sistema irá segmentar os dados existentes em clusters, que serão definidos tendo por base a categoria indicada, e os ratings superiores a 8. Assim sendo este sistema comunica com o utilizador por uma interface gráfica, onde é fornecida pelo utilizador a sua categoria preferida. Posteriormente esta indicação será utilizada para indicar por via dos dos wrapped nodes *Top Directors*, *Top Movies per Director* e *Top Movies*, os diretores com melhor rating, os melhores filmes por director e o top 5 filmes, respetivamente.

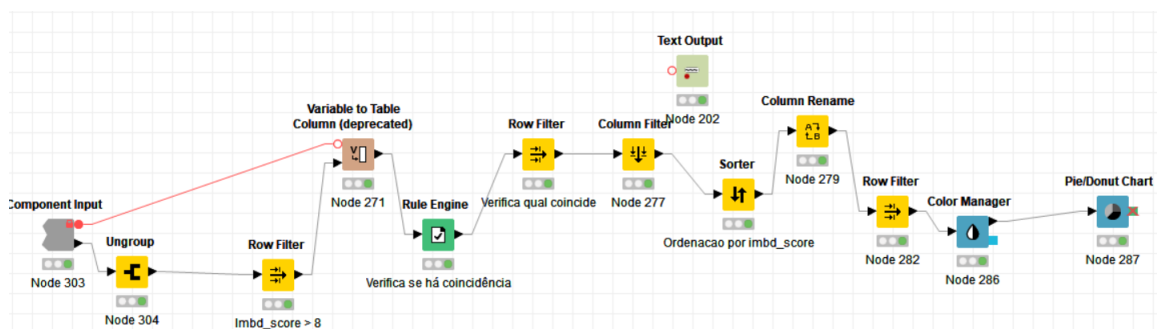


Figura 10: Top 5 Filmes

Na figura 10, demonstramos como o output é determinado, de notar que esta forma é aplicada da mesma forma para determinar os diretores com melhor rating e os melhores filmes por director, mudando unicamente o nodo final que apresenta os resultados graficamente.

Tal como indicado, utilizamos clusters para obter o mesmo tipo de filmes existentes nesses mesmos. Como foi referido anteriormente, este sistema é baseado na segmentação dos dados de forma a produzir recomendações, pelo que a sua otimização seja importante de forma a criar sugestões com melhor qualidade. Assim sendo fazemos uma primeira filtração de filmes que não possuem um rating igual ou superior a 8, após isso é verificado qual dos filmes é do género mencionado, para posteriormente os que não são serem removidos.

5.2 Baseado em Regras Associativas

Com a existência de um volume grande dados, é possível desenvolver um sistema baseado em comportamentos e tendências observados nesses dados. Assim é possível fazer recomendações de filmes, tendo como referência esses mesmos comportamentos. Tal como as recomendações enunciadas anteriormente, primeiro de tudo é requerido ao utilizador seleccionar por exemplo o seu filme, director ou género preferido, tudo isto via os wrapped metanodes, "*Movie Selection*", "*Director Selection*" e "*Genre Selection*",

respetivamente. Posteriormente este input será processado com as geradas previamente.

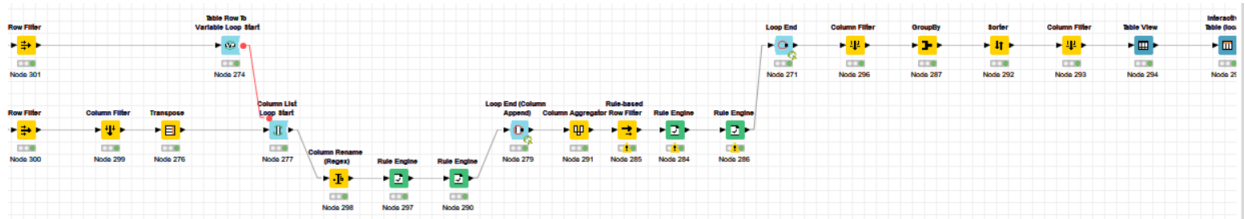


Figura 11: Comparação das regras

Nesta figura apresentada em cima, é mostrado o processo responsável pela comparação das regras associativas, este mesmo processo é aplicado nos wrapped metanodes "*AR_Movie_Output*", "*AR_Genre_Output*". O objetivo passa por obter o consequente que melhor se aplica ao input dado, sendo o consequente o filme/s ou gênero/s que serão recomendados, para isso passa por diversas verificações, as quais serão removidas todas aquelas que falharem essas mesmas. Por outras palavras, é testado se um filme ou gênero esta presenta nos antecedentes, para assim se executar uma filtragem das associações.

Para o wrapped metanode "*AR_Director_Output*", utiliza dois processos, um deles com as mesma linha de pensamento mas adaptado as suas circunstâncias.

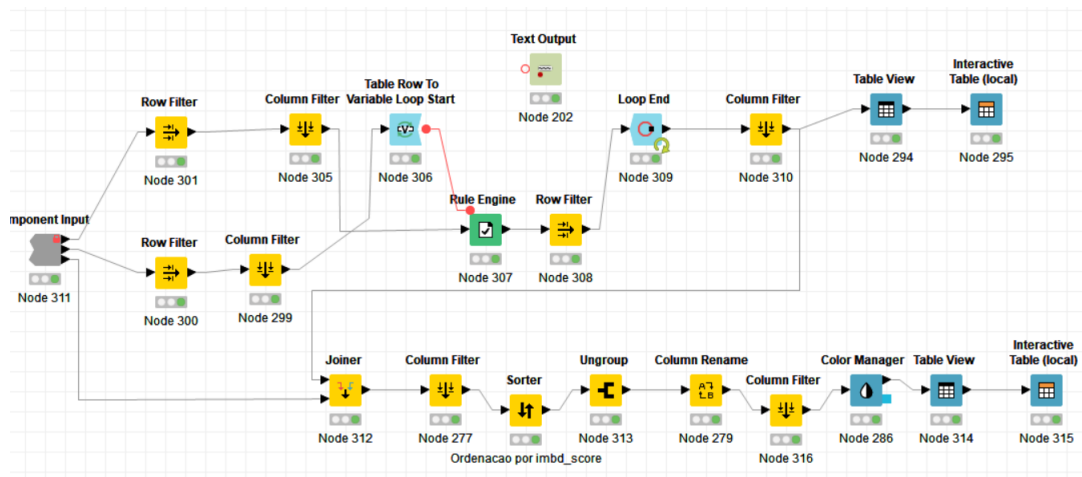
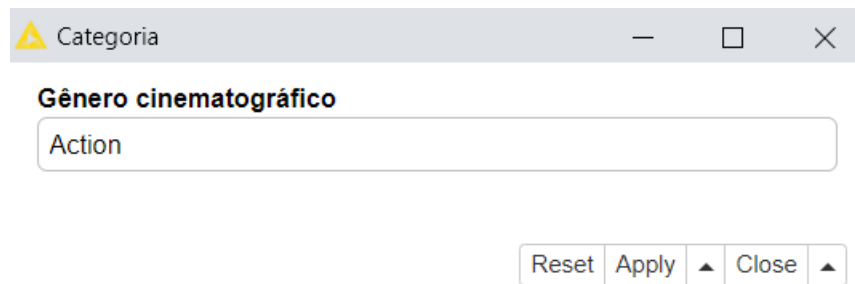


Figura 12: Processo para sugestão de directores

O outro processo utiliza o output gerado no processo de cima, para criar uma lista dos filmes desses mesmos directores, que posteriormente será apresentada ao utilizador.

6 Resultados Obtidos

6.1 Baseado me Clusters



A screenshot of a web application window titled "Categoria". Inside the window, there is a label "Gênero cinematográfico" above a text input field containing the word "Action". Below the input field, there are four buttons: "Reset", "Apply", "Close", and a small upward-pointing triangle button.

Figura 13: Interface gráfica responsável por obter a categoria preferida do utilizador

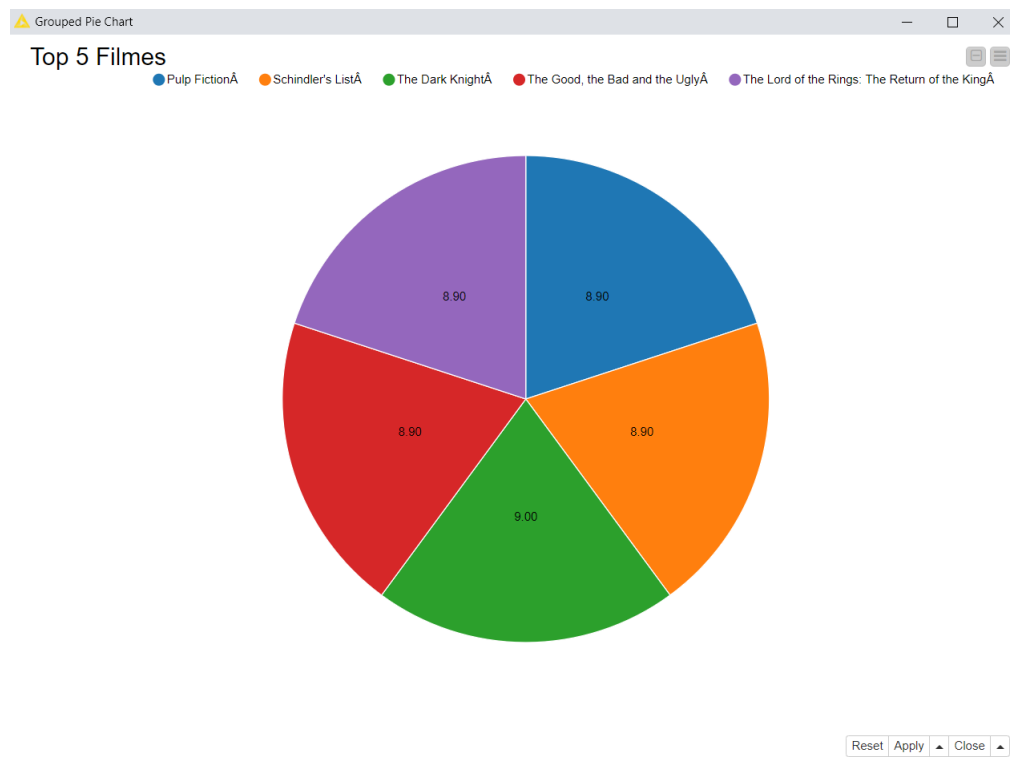


Figura 14: Top 5 filmes, juntamente com o seu rating

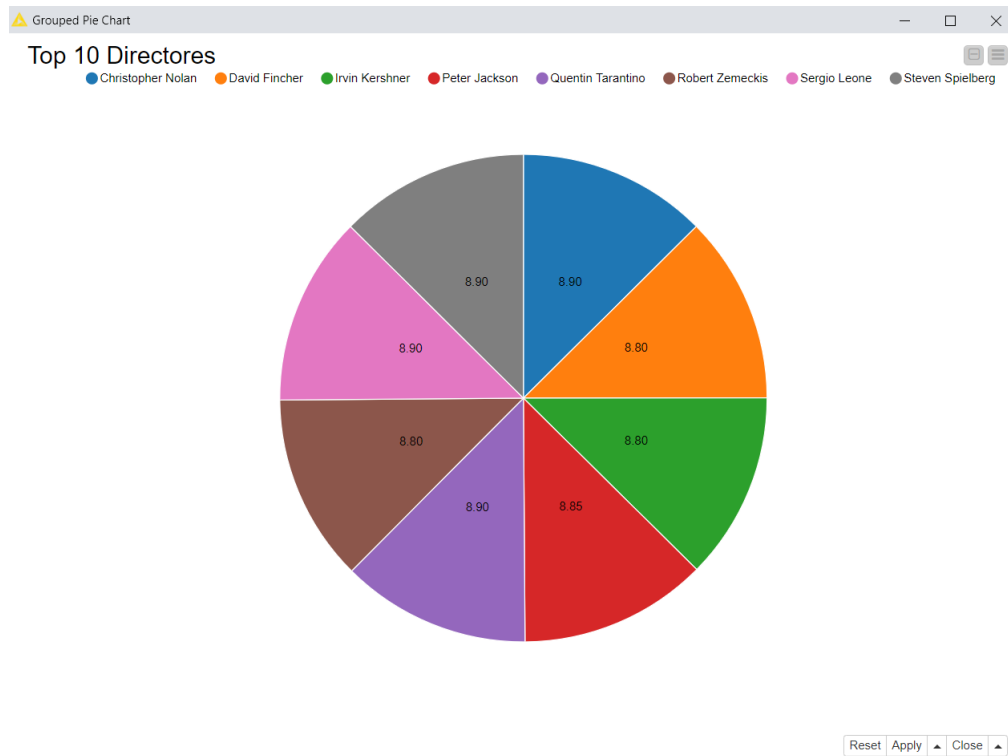


Figura 15: Top directores, juntamente com o seu rating médio

JavaScript Table View

Top Filmes por Director

Show 10 entries Search:

	Director	Rating	Título
	Christopher Nolan	9	The Dark Knight
	Peter Jackson	8.9	The Lord of the Rings: The Return of the King
	Quentin Tarantino	8.9	Pulp Fiction
	Sergio Leone	8.9	The Good, the Bad and the Ugly
	Steven Spielberg	8.9	Schindler's List
	Christopher Nolan	8.8	Inception
	David Fincher	8.8	Fight Club
	Irvin Kershner	8.8	Star Wars: Episode V - The Empire Strikes Back
	Peter Jackson	8.8	The Lord of the Rings: The Fellowship of the Ring
	Robert Zemeckis	8.8	Forrest Gump

Showing 1 to 10 of 10 entries Previous 1 Next

Reset Apply Close

Figura 16: Listagem dos melhores filmes, dos directores recomendados

6.2 Baseado em Regras Associativas

Movie Selection

Label

Association Rules Recommender System

Selecione alguns filmes que já assistiu

Seleção de Filmes

Show 10 entries Search: Star W

movie_title
<input type="checkbox"/> Star Trek II: The Wrath of Khan
<input type="checkbox"/> Star Wars: Episode I - The Phantom Menace
<input type="checkbox"/> Star Wars: Episode II - Attack of the Clones
<input checked="" type="checkbox"/> Star Wars: Episode III - Revenge of the Sith
<input checked="" type="checkbox"/> Star Wars: Episode IV - A New Hope
<input checked="" type="checkbox"/> Star Wars: Episode V - The Empire Strikes Back
<input type="checkbox"/> Star Wars: Episode VI - Return of the Jedi
<input type="checkbox"/> The Men Who Stare at Goats

Showing 1 to 8 of 8 entries (filtered from 4,468 total entries)

Previous 1 Next

Figura 17: Escolha dos filmes

Genre Selection

Seleção de Género

Show 10 entries Search: biogra

genres
<input type="checkbox"/> Biography Drama History Thriller War
<input type="checkbox"/> Biography Drama History War
<input type="checkbox"/> Biography Drama Music
<input checked="" type="checkbox"/> Biography Drama Music Musical
<input checked="" type="checkbox"/> Biography Drama Music Romance
<input checked="" type="checkbox"/> Biography Drama Romance
<input checked="" type="checkbox"/> Biography Drama Romance Sport
<input checked="" type="checkbox"/> Biography Drama Romance War
<input type="checkbox"/> Biography Drama Romance Western
<input type="checkbox"/> Biography Drama Sport

Showing 81 to 90 of 94 entries (filtered from 914 total entries)

Previous 1 ... 6 7 8 9 10 Next

Figura 18: Escolha dos géneros

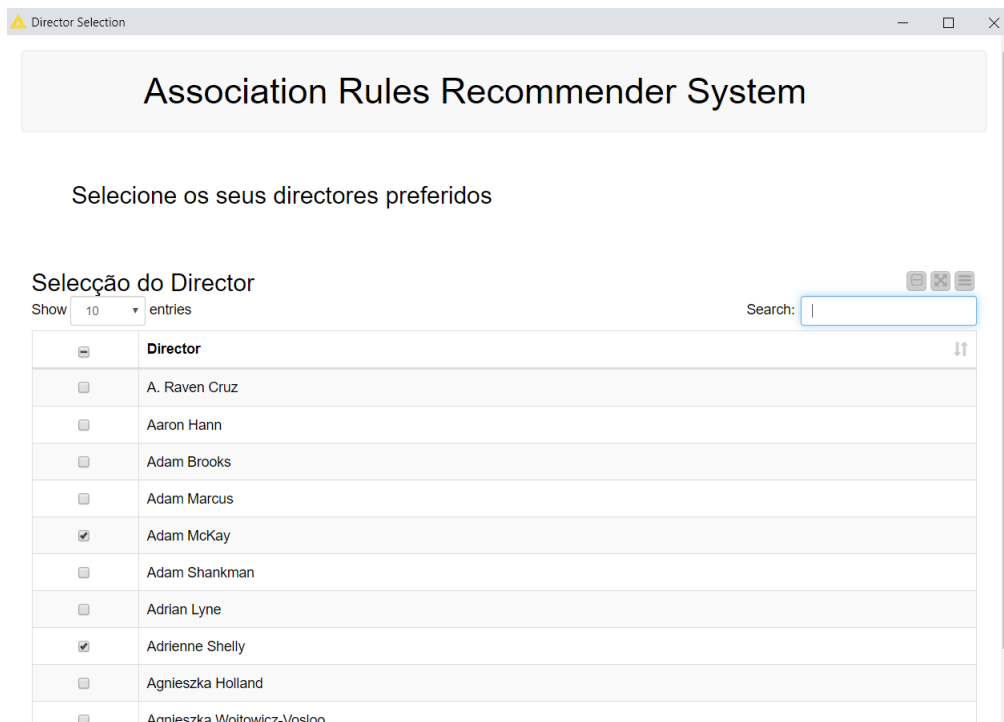


Figura 19: Escolha dos directores

De seguida apresentamos de cada uma das selecções respetivamente.

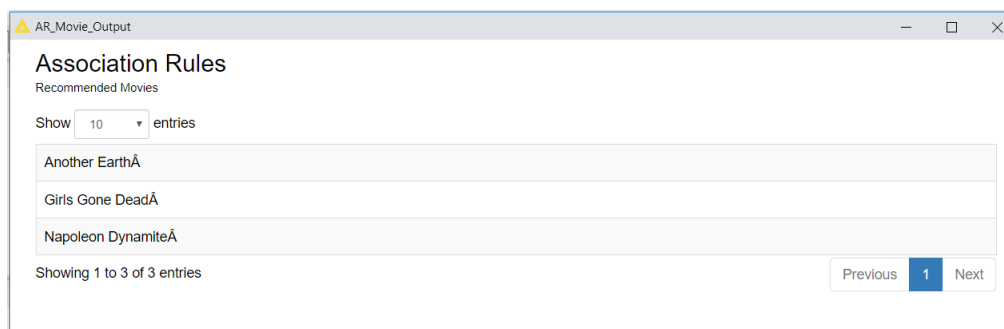


Figura 20: Filmes recomendados pelas regras

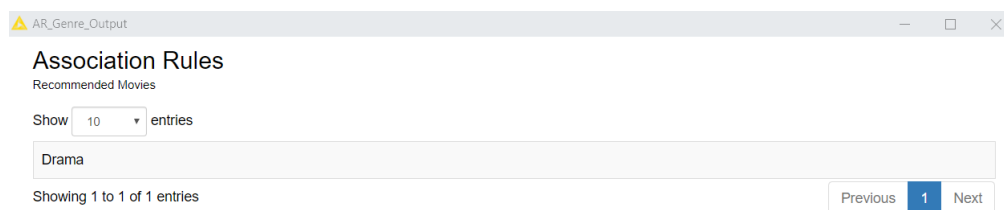


Figura 21: Géneros recomendados pelas regras

Melhores Diretores Relacionados

Association Rules
Recommended Movies

Show entries

Tim Burton
Steven Spielberg
Chris Columbus
Robert Rodriguez
Tarsem Singh
Luc Besson
Richard Donner
Jon Favreau
Tim Burton
Robert Rodriguez

Showing 1 to 10 of 15 entries

Previous **1** 2 Next

Figura 22: Directores recomendados pelas regras

Esta página para além de apresentar os directores recomendáveis pelas regras associativas, também apresenta os seus melhores filmes.

Association Rules
Recommended Movies

Show entries

Steven Spielberg	Indiana Jones and the Kingdom of the Crystal Skull
Steven Spielberg	The BFG
Steven Spielberg	War of the Worlds
Steven Spielberg	The Adventures of Tintin
Steven Spielberg	Minority Report
Steven Spielberg	A.I. Artificial Intelligence
Steven Spielberg	The Lost World: Jurassic Park
Steven Spielberg	The Terminal
Steven Spielberg	Munich
Steven Spielberg	Hook

Showing 1 to 10 of 186 entries

Previous **1** 2 3 4 5 ... 19 Next

Figura 23: Filmes dos directores recomendados

7 Sugestões e Recomendações

Necessidade de melhorar o sistema com objetivo de obter recomendações mais precisas, o que implicaria uma otimização na arquitectura.

Relativamente as regras de associação, as escolhas poderiam ser mais poderadas, para as configurações, tais como suporte e confiança mínima, por outro lado a execução será mais demorada.

Por fim consideramos que um dataset mais completo poderia trazer melhorias as implementações já existentes e traria novas implementações, como por exemplo filtragem

colaborativa, que não pode ser aplicada pois não utilizamos um dataset contendo dados de utilizadores.

8 Conclusão

Concluída a realização deste trabalho, permitiu-nos adequir sobre o funcionamento de um sistema de recomendação.

Este mesmo sistema, permitiu-nos perceber a real importância de um dataset completo de forma a obter-mos um sistema recomendação preciso e robusto.

Por fim e com a conclusão deste trabalho, apercebemo-nos mais uma vez das capacidades que a plataforma utilizada, *Knime*, nos permite, ficando também uma certa curiosidade em utilizar outros algoritmos de *Machine Learning* nela mesma.